BEST AVAILABLE IMAGES

Defective images within this document are accurate representations of the original documents submitted by the applicant.

Defects in the images may include (but are not limited to):

- BLACK BORDERS
- TEXT CUT OFF AT TOP, BOTTOM OR SIDES
- FADED TEXT
- ILLEGIBLE TEXT
- SKEWED/SLANTED IMAGES
- COLORED PHOTOS
- BLACK OR VERY BLACK AND WHITE DARK PHOTOS
- GRAY SCALE DOCUMENTS

IMAGES ARE BEST AVAILABLE COPY.

As rescanning documents *will not* correct images, please do not report the images to the Image Problem Mailbox.

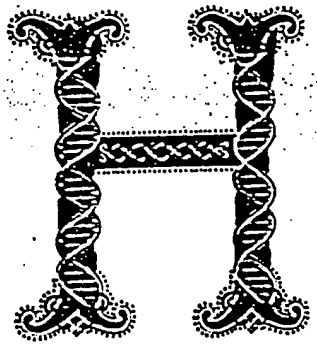
characterize the protein. A starting material that can only be used to produce a final product does not have a substantial asserted utility in those instances where the final product is not supported by a specific and substantial utility. In this case none of the proteins that are to be produced as final products resulting from processes involving the claimed cDNA have asserted or identified specific and substantial utilities. The research contemplated by Applicants to characterize potential protein products, especially their biological activities, does not constitute a specific and substantial utility. Identifying and studying the properties of the protein itself or the mechanisms in which the protein is involved does not define a "real world" context of use. Note, because the claimed invention is not supported by a specific and substantial asserted utility for the reasons set forth above, credibility has not been assessed. Neither the specification as filed nor any art of record discloses or suggests any property or activity for the cDNA compounds such that another non-asserted utility would be well established for the compounds.

Claim 1 is also rejected under 35 U.S.C. § 112, first paragraph. Specifically, since the claimed invention is not supported by either a specific and substantial asserted utility or a well established utility for the reasons set forth above, one skilled in the art would not know how to use the claimed invention.

Example 10: <u>DNA Fragment encoding a Full Open Reading Frame (ORF)</u>

Specification: The specification discloses that a cDNA library was prepared from human kidney epithelial cells and 5000 members of this library were

THE HUMAN GENOME



Humanity has been given a great gift. With the completion of the human genome sequence, we have received a powerful tool for unlocking the secrets of our genetic heritage and for finding our place among the other participants in the adventure of life.

This week's issue of *Science* contains the report of the sequencing of the human genome from a group of authors led by Craig Venter of Celera Genomics. The report of the sequencing of the human genome from the publicly funded consortium of laboratories led by Francis Collins appears in this week's *Nature*. This stunning achievement has been portrayed—often unfairly—as a competition between two ventures, one public and one private. That characterization detracts from the awesome accomplishment jointly unveiled this week. In truth, each project contributed to the other. The inspired vision that launched the publicly funded project roughly 10 years ago reflected, and now rewards, the confidence of those who believe that the pursuit of large-scale fundamental problems in the life sciences is in the national interest. The technical innovation and drive of Craig Venter and his colleagues made it possible to celebrate this accomplishment far sooner than was believed possible. Thus, we can salute what has become, in the end, not a contest but a marriage (perhaps encouraged by shotgun) between public funding and private entrepreneurship.

A historic moment for the scientific endeavor.

There are excellent scientific reasons for applauding an outcome that has given us two winners. Two sequences are better than one; the opportunity for comparison and convergence is invaluable. Indeed, a real-world proof of the importance of access to both sets of data can be found in the pages of this issue of *Science*, in the comparative analysis by Olivier *et al.* (p. 1298).

Although we have made the point before, it is worth repeating that the sequencing of the human genome represents, not an ending, but the beginning of a new approach to biology. As Galas says in his Viewpoint (p. 1257), the knowledge that all of the genetic components of any process can be identified will give extraordinary new power to scientists. Because of this breakthrough, research can evolve from analyzing the effects of individual genes to a more integrated view that examines whole ensembles of genes as they interact to form a living human being. Several articles in this issue highlight how this approach is already beginning to revolutionize the way we look at human disease.

This has been a massive project, on a scale unparalleled in the history of biology, but of course it has built on the scientific insights of centuries of investigators. By coincidence, this landmark announcement falls during the week of the anniversary of the birth of Charles Darwin. Darwin's message that the survival of a species can depend on its ability to evolve in the face of change is peculiarly pertinent to discussions that have gone on in the past year over access to the Celera data. (Full information regarding the agreements that were reached to make the data available can be found at www.sciencemag.org/feature/data/announcement/gsp.shl.) We are willing to be flexible in allowing data repositories other than the traditional GenBank, while insisting on access to all the data needed to verify conclusions. In this domain, change is everywhere: Commercial researchers are producing more and more potentially valuable sequences, yet (at least in the United States) laws governing databases provide scant protection against piracy. Had the Celera data been kept secret, it would have been a serious loss to the scientific community. We hope that our adaptability in the face of change will enable other proprietary data to be published after peer review, in a way that satisfies our continuing commitment to full access.

It should be no surprise that an achievement so stunning, and so carefully watched, has created new challenges for the scientific venture. *Science* is proud to have played a role in bringing this discovery onto the public stage. It is literally true that this is a historic moment for the scientific endeavor. The human genome has been called the Book of Life. Rather, it is a library, in which, with rules that encourage exploration and reward creativity, we can find many of the books that will help define us and our place in the great tapestry of life.

Barbara R. Jasny and Donald Kennedy

sequenced and open reading frames were identified. The specification discloses a Table that indicates that one member of the library having SEQ ID NO: 2 has a high level of homology to a DNA ligase. The specification teaches that this complete ORF (SEQ ID NO: 2) encodes SEQ ID NO: 3. An alignment of SEQ ID NO: 3 with known amino acid sequences of DNA ligases indicates that there is a high level of sequence conservation between the various known ligases. The overall level of sequence similarity between SEQ ID NO: 3 and the consensus sequence of the known DNA ligases that are presented in the specification reveals a similarity score of 95%. A search of the prior art confirms that SEQ ID NO: 2 has high homology to DNA Ligase encoding nucleic acids and that the next highest level of homology is to alpha-actin. However, the latter homology is only 50%. Based on the sequence homologies, the specification asserts that SEQ ID NO: 2 encodes a DNA ligase.

Claim 1: An isolated and purified nucleic acid comprising SEQ ID NO: 2.

Analysis: The following analysis includes the questions that need to be asked according to the guidelines and the answers to those questions based on the above facts:

1) Based on the record, is there a "well established utility" for the claimed invention? Based upon applicant's disclosure and the results of the PTO search, there is no reason to doubt the assertion that SEQ ID NO: 2 encodes a DNA ligase. Further, DNA ligases have a well-established use in the molecular biology art based on this class of protein's ability to ligate DNA. Consequently the answer to the question is yes.

Note that if there is a well-established utility already associated with the claimed invention, the utility need not be asserted in the specification as filed. In order to determine whether the claimed invention has a well-established utility the examiner must determine that the invention has a specific, substantial and credible utility that would have been readily apparent to one of skill in the art. In this case SEQ ID NO: 2 was shown to encode a DNA ligase that the artisan would have recognized as having a specific, substantial and credible utility based on its enzymatic activity.

Thus, the conclusion reached from this analysis is that a 35 U.S.C. § 101 rejection and a 35 U.S.C. § 112, first paragraph, utility rejection should not be made.

Example 11: Animals with Uncharacterized Human Genes

Specification: Kidney cells from a patient with Polycystic Kidney (PCK) Disease have been used to make a cDNA library. From this library 8000 nucleotide "fragments" have been sequenced but not yet used to express proteins in a transformed host cell nor have they been characterized in any other way. The 50 longest fragments, SEQ ID NO: 1-50, respectively, have been used to make transgenic mice. None of the 50 lines of mice have developed Polycystic Kidney Disease to date. The asserted utility is the use of the mice to research human genes from diseased human kidneys. The disease is inheritable, but chromosomal loci have not yet been identified. Neither the absence or presence of a specific protein has been identified with the disease condition.

>AF410459 TRANSLATION:AF410459 NID: gi 19071208 gb AF410459.1 Homo sapiens CD109 (CD109) mRNA, complete cds
Length = 5883

Score = 2854 bits (7317), Expect = 0.0
Identities = 1427/1445 (98%), Positives = 1428/1445 (98%), Gaps = 17/1445 (1%)
Frame = +2

```
Query: 1    MQGPPLLTAAHLLCVCTAALAVAPGPRFLVTAPGIIRPGGNVTIGVELLEHCPSQVTVKA 60
            MQGPPLLTAAHLLCVCTAALAVAPGPRFLVTAPGIIRPGGNVTIGVELLEHCPSQVTVKA
Sbjct: 113  MQGPPLLTAAHLLCVCTAALAVAPGPRFLVTAPGIIRPGGNVTIGVELLEHCPSQVTVKA 292

Query: 61   ELLKTASNLTVSVLEAEGVFEKGSFKTLTLPSLPLNSADEIYELRVTGRTQDEILFSNST 120
            ELLKTASNLTVSVLEAEGVFEKGSFKTLTLPSLPLNSADEIYELRVTGRTQDEILFSNST
Sbjct: 293  ELLKTASNLTVSVLEAEGVFEKGSFKTLTLPSLPLNSADEIYELRVTGRTQDEILFSNST 472

Query: 121  RLSFETKRISVFIQTDKALYKPKQEVKFRIVTLFSDFKPYKTSLNILIKDPKSNLIQQWL 180
            RLSFETKRISVFIQTDKALYKPKQEVKFRIVTLFSDFKPYKTSLNILIKDPKSNLIQQWL
Sbjct: 473  RLSFETKRISVFIQTDKALYKPKQEVKFRIVTLFSDFKPYKTSLNILIKDPKSNLIQQWL 652

Query: 181  SQQSDLGVISKTFQLSSHPILGDWSIQVQVNDQTYYQSFQVSEYVLPKFEVTLQTPLYCS 240
            SQQSDLGVISKTFQLSSHPILGDWSIQVQVNDQTYYQSFQVSEYVLPKFEVTLQTPLYCS
Sbjct: 653  SQQSDLGVISKTFQLSSHPILGDWSIQVQVNDQTYYQSFQVSEYVLPKFEVTLQTPLYCS 832

Query: 241  MNSKHLNGTITAKYTYGKPVKGDVTLTFLPLSFWGKKKNITKTFKINGSANFSFNDEEMK 300
            MNSKHLNGTITAKYTYGKPVKGDVTLTFLPLSFWGKKKNITKTFKINGSANFSFNDEEMK
Sbjct: 833  MNSKHLNGTITAKYTYGKPVKGDVTLTFLPLSFWGKKKNITKTFKINGSANFSFNDEEMK 1012

Query: 301  NVMDSSNGLSEYLDLSSPGPVEILTTVTESVTGISRNVSTNVFFKQHDYIIEFFDYTTVL 360
            NVMDSSNGLSEYLDLSSPGPVEILTTVTESVTGISRNVSTNVFFKQHDYIIEFFDYTTVL
Sbjct: 1013 NVMDSSNGLSEYLDLSSPGPVEILTTVTESVTGISRNVSTNVFFKQHDYIIEFFDYTTVL 1192

Query: 361  KPSLNFTATVKVTRADGNQLTLEERRNNVVITVTQRNYTEYWSGSNSGNQKMEAVQKINY 420
            KPSLNFTATVKVTRADGNQLTLEERRNNVVITVTQRNYTEYWSGSNSGNQKMEAVQKINY
Sbjct: 1193 KPSLNFTATVKVTRADGNQLTLEERRNNVVITVTQRNYTEYWSGSNSGNQKMEAVQKINY 1372

Query: 421  TVPQSGTFKIEFPILEDSSELQLKAYFLGSKSSMAVHSLFKSPSKTYIQLKTRDENIKVG 480
            TVPQSGTFKIEFPILEDSSELQLKAYFLGSKSSMAVHSLFKSPSKTYIQLKTRDENIKVG
Sbjct: 1373 TVPQSGTFKIEFPILEDSSELQLKAYFLGSKSSMAVHSLFKSPSKTYIQLKTRDENIKVG 1552

Query: 481  SPFELVVSGNKRLKELSYMVVSRGQLVAVGKQNSTMFSLTPENSWTPKACVIVYYIEDDG 540
            SPFELVVSGNKRLKELSYMVVSRGQLVAVGKQNSTMFSLTPENSWTPKACVIVYYIEDDG
Sbjct: 1553 SPFELVVSGNKRLKELSYMVVSRGQLVAVGKQNSTMFSLTPENSWTPKACVIVYYIEDDG 1732

Query: 541  EIISDVLKIPVQLVFKNKIKLYWSKVKAEPSEKVSLRISVTQPDSIVGIVAVDKSVNLMN 600
            EIISDVLKIPVQLVFKNKIKLYWSKVKAEPSEKVSLRISVTQPDSIVGIVAVDKSVNLMN
Sbjct: 1733 EIISDVLKIPVQLVFKNKIKLYWSKVKAEPSEKVSLRISVTQPDSIVGIVAVDKSVNLMN 1912

Query: 601  ASNDITMENVVHELELYNTGYYLGMFMNSFAVFQECGLWVLTDANLTKDYIDGVYDNAEY 660
            ASNDITMENVVHELELYNTGYYLGMFMNSFAVFQECGLWVLTDANLTKDYIDGVYDNAEY
Sbjct: 1913 ASNDITMENVVHELELYNTGYYLGMFMNSFAVFQECGLWVLTDANLTKDYIDGVYDNAEY 2092

Query: 661  AERFMEENEGHIVDIHDFSLGSSPHVRKHFPETWIWLDTNMGYRIYQEFEVTVPDSITSW 720
            AERFMEENEGHIVDIHDFSLGSSPHVRKHFPETWIWLDTNMGYRIYQEFEVTVPDSITSW
Sbjct: 2093 AERFMEENEGHIVDIHDFSLGSSPHVRKHFPETWIWLDTNMGYRIYQEFEVTVPDSITSW 2272
```

```
Query: 721  VATGFVISEDLGLGLTTTPVELQAFQPFFIFLNLPYSVIRGEEFALEITIFNYLKDATEV 780
            VATGFVISEDLGLGLTTTPVELQAFQPFFIFLNLPYSVIRGEEFALEITIFNYLKDATEV
Sbjct: 2273 VATGFVISEDLGLGLTTTPVELQAFQPFFIFLNLPYSVIRGEEFALEITIFNYLKDATEV 2452

Query: 781  KVIIEKSDKFDILMTSSEINATGHQQTLLVPSEDGATVLFPIRPTHLGEIPITVTALSPT 840
            KVIIEKSDKFDILMTS+EINATGHQQTLLVPSEDGATVLFPIRPTHLGEIPITVTALSPT
Sbjct: 2453 KVIIEKSDKFDILMTSNEINATGHQQTLLVPSEDGATVLFPIRPTHLGEIPITVTALSPT 2632

Query: 841  ASDAVTQMILVKAEGIEKSYSQSILLDLTDNRLQSTLKTLSFSFPPNTVTGSERVQITAI 900
            ASDAVTQMILVKAEGIEKSYSQSILLDLTDNRLQSTLKTLSFSFPPNTVTGSERVQITAI
Sbjct: 2633 ASDAVTQMILVKAEGIEKSYSQSILLDLTDNRLQSTLKTLSFSFPPNTVTGSERVQITAI 2812

Query: 901  GDVLGPSINGLASLIRMPYGCGEQNMINFAPNIYILDYLTKKKQLTDNLKEKALSFMRQG 960
            GDVLGPSINGLASLIRMPYGCGEQNMINFAPNIYILDYLTKKKQLTDNLKEKALSFMRQG
Sbjct: 2813 GDVLGPSINGLASLIRMPYGCGEQNMINFAPNIYILDYLTKKKQLTDNLKEKALSFMRQG 2992

Query: 961  YQRELLYQREDGSFSAFGNYDPSGSTWLSAFVLRCFLEADPYIDIDQNVLHRTYTWLKGH 1020
            YQRELLYQREDGSFSAFGNYDPSGSTWLSAFVLRCFLEADPYIDIDQNVLHRTYTWLKGH
Sbjct: 2993 YQRELLYQREDGSFSAFGNYDPSGSTWLSAFVLRCFLEADPYIDIDQNVLHRTYTWLKGH 3172

Query: 1021 QKSNGEFWDPGRVIHSELQGGNKSPVTLTAYIVTSLLGYRKYQPNIDVQESIHFLESEFS 1080
            QKSNGEFWDPGRVIHSELQGGNKSPVTLTAYIVTSLLGYRKYQPNIDVQESIHFLESEFS
Sbjct: 3173 QKSNGEFWDPGRVIHSELQGGNKSPVTLTAYIVTSLLGYRKYQPNIDVQESIHFLESEFS 3352

Query: 1081 RGISDNYTLALITYALSSVGSPKAKEALNMLTWRAEQEGGMQFWVSSESKLSDSWQPRSL 1140
            RGISDNYTLALITYALSSVGSPKAKEALNMLTWRAEQEGGMQFWVSSESKLSDSWQPRSL
Sbjct: 3353 RGISDNYTLALITYALSSVGSPKAKEALNMLTWRAEQEGGMQFWVSSESKLSDSWQPRSL 3532

Query: 1141 DIEVAAYALLSHFLQFQTSEGIPIMRWLSRQRNSLGGFASTQDTTVALKALSEFAALMNT 1200
            DIEVAAYALLSHFLQFQTSEGIPIMRWLSRQRNSLGGFASTQDTTVALKALSEFAALMNT
Sbjct: 3533 DIEVAAYALLSHFLQFQTSEGIPIMRWLSRQRNSLGGFASTQDTTVALKALSEFAALMNT 3712

Query: 1201 ERTNIQVTVTGPSSPSP----------------LAVVQPMAVNISANGFGFAICQLNVV 1243
            ERTNIQVTVTGPSSPSP                LAVVQPMAVNISANGFGFAICQLNVV
Sbjct: 3713 ERTNIQVTVTGPSSPSPVKFLIDTHNRLLLQTAELAVVQPMAVNISANGFGFAICQLNVV 3892

Query: 1244 YNVKASGSSRRRRSIQNQEAFDLDVAVKENKDDLNHVDLNVCTSFSGPGRSGMALMEVNL 1303
            YNVKASGSSRRRRSIQNQEAFDLDVAVKENKDDLNHVDLNVCTSFSGPGRSGMALMEVNL
Sbjct: 3893 YNVKASGSSRRRRSIQNQEAFDLDVAVKENKDDLNHVDLNVCTSFSGPGRSGMALMEVNL 4072

Query: 1304 LSGFMVPSEAISLSETVKKVEYDHGKLNLYLDSVNETQFCVNIPAVRNFKVSNTQDASVS 1363
            LSGFMVPSEAISLSETVKKVEYDHGKLNLYLDSVNETQFCVNIPAVRNFKVSNTQDASVS
Sbjct: 4073 LSGFMVPSEAISLSETVKKVEYDHGKLNLYLDSVNETQFCVNIPAVRNFKVSNTQDASVS 4252

Query: 1364 IVDYYEPRRQAVRSYNSEVKLSSCDLCSDVQGCRPCEDGASGSHHHSSVIFIFCFKLLYF 1423
            IVDYYEPRRQAVRSYNSEVKLSSCDLCSDVQGCRPCEDGASGSHHHSSVIFIFCFKLLYF
Sbjct: 4253 IVDYYEPRRQAVRSYNSEVKLSSCDLCSDVQGCRPCEDGASGSHHHSSVIFIFCFKLLYF 4432

Query: 1424 MELWL 1428
            MELWL
Sbjct: 4433 MELWL 4447
```

1: AF410459. Homo sapiens CD10...[gi:19071208]

```
LOCUS       AF410459                5883 bp    mRNA    linear   PRI 02-MAR-2002
DEFINITION  Homo sapiens CD109 (CD109) mRNA, complete cds.
ACCESSION   AF410459
VERSION     AF410459.1  GI:19071208
KEYWORDS    .
SOURCE      Homo sapiens (human)
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.
REFERENCE   1  (bases 1 to 5883)
  AUTHORS   Lin,M., Sutherland,D.R., Horsfall,W., Totty,N., Yeo,E., Nayar,R.,
            Wu,X.F. and Schuh,A.C.
  TITLE     Cell surface antigen CD109 is a novel member of the alpha(2)
            macroglobulin/C3, C4, C5 family of thioester-containing proteins
  JOURNAL   Blood 99 (5), 1683-1691 (2002)
  MEDLINE   21849742
   PUBMED   11861284
REFERENCE   2  (bases 1 to 5883)
  AUTHORS   Lin,M., Sutherland,D.R., Horsfall,W., Totty,N., Yeo,E., Nayar,R.,
            Wu,X.-F. and Schuh,A.C.
  TITLE     Direct Submission
  JOURNAL   Submitted (14-AUG-2001) Medicine, University of Toronto, 1 King's
            College Circle, Room 7366, Toronto, Ontario M5S 1A8, Canada
FEATURES             Location/Qualifiers
     source          1..5883
                     /organism="Homo sapiens"
                     /mol_type="mRNA"
                     /db_xref="taxon:9606"
                     /clone="K1"
     gene            1..5883
                     /gene="CD109"
     5'UTR           1..112
                     /gene="CD109"
     CDS             113..4450
                     /gene="CD109"
                     /note="associated with the Gov alloantigen system"
                     /codon_start=1
                     /product="CD109"
                     /protein_id="AAL84159.1"
                     /db_xref="GI:19071209"
                     /translation="MQGPPLLTAAHLLCVCTAALAVAPGPRFLVTAPGIIRPGGNVTI
                     GVELLEHCPSQVTVKAELLKTASNLTVSVLEAEGVFEKGSFKTLTLPSLPLNSADEIY
                     ELRVTGRTQDEILFSNSTRLSFETKRISVFIQTDKALYKPKQEVKFRIVTLFSDFKPY
                     KTSLNILIKDPKSNLIQQWLSQQSDLGVISKTFQLSSHPILGDWSIQVQVNDQTYYQS
                     FQVSEYVLPKFEVTLQTPLYCSMNSKHLNGTITAKYTYGKPVKGDVTLTFLPLSFWGK
                     KKNITKTFKINGSANFSFNDEEMKNVMDSSNGLSEYLDLSSPGPVEILTTVTESVTGI
                     SRNVSTNVFFKQHDYIIEFFDYTTVLKPSLNFTATVKVTRADGNQLTLEERRNNVVIT
                     VTQRNYTEYWSGSNSGNQKMEAVQKINYTVPQSGTFKIEFPILEDSSELQLKAYFLGS
                     KSSMAVHSLFKSPSKTYIQLKTRDENIKVGSPFELVVSGNKRLKELSYMVVSRGQLVA
```

```
                     VGKQNSTMFSLTPENSWTPKACVIVYYIEDDGEIISDVLKIPVQLVFKNKIKLYWSKV
                     KAEPSEKVSLRISVTQPDSIVGIVAVDKSVNLMNASNDITMENVVHELELYNTGYYLG
                     MFMNSFAVFQECGLWVLTDANLTKDYIDGVYDNAEYAERFMEENEGHIVDIHDFSLGS
                     SPHVRKHFPETWIWLDTNMGYRIYQEFEVTVPDSITSWVATGFVISEDLGLGLTTTPV
                     ELQAFQPFFIFLNLPYSVIRGEEFALEITIFNYLKDATEVKVIIEKSDKFDILMTSNE
                     INATGHQQTLLVPSEDGATVLFPIRPTHLGEIPITVTALSPTASDAVTQMILVKAEGI
                     EKSYSQSILLDLTDNRLQSTLKTLSFSFPPNTVTGSERVQITAIGDVLGPSINGLASL
                     IRMPYGCGEQNMINFAPNIYILDYLTKKKQLTDNLKEKALSFMRQGYQRELLYQREDG
                     SFSAFGNYDPSGSTWLSAFVLRCFLEADPYIDIDQNVLHRTYTWLKGHQKSNGEFWDP
                     GRVIHSELQGGNKSPVTLTAYIVTSLLGYRKYQPNIDVQESIHFLESEFSRGISDNYT
                     LALITYALSSVGSPKAKEALNMLTWRAEQEGGMQFWVSSESKLSDSWQPRSLDIEVAA
                     YALLSHFLQFQTSEGIPIMRWLSRQRNSLGGFASTQDTTVALKALSEFAALMNTERTN
                     IQVTVTGPSSPSPVKFLIDTHNRLLLQTAELAVVQPMAVNISANGFGFAICQLNVVYN
                     VKASGSSRRRRSIQNQEAFDLDVAVKENKDDLNHVDLNVCTSFSGPGRSGMALMEVNL
                     LSGFMVPSEAISLSETVKKVEYDHGKLNLYLDSVNETQFCVNIPAVRNFKVSNTQDAS
                     VSIVDYYEPRRQAVRSYNSEVKLSSCDLCSDVQGCRPCEDGASGSHHHSSVIFIFCFK
                     LLYFMELWL"
     3'UTR           4451..5883
                     /gene="CD109"
     misc_feature    4748
                     /gene="CD109"
                     /note="alternative polyA site found on clone K1"
ORIGIN
        1 ctaaactcga attaagaggg aaaaaaaatc agggaggagg tggcaagcca caccccacgg
       61 tgcccgcgaa cttccccggc agcggactgt agcccaggca gacgccgtcg agatgcaggg
      121 cccaccgctc ctgaccgccg cccacctcct ctgcgtgtgc accgccgcgc tggccgtggc
      181 tcccgggcct cggtttctgg tgacagcccc agggatcatc aggcccggag gaaatgtgac
      241 tattggggtg gagcttctgg aacactgccc ttcacaggtg actgtgaagg cggagctgct
      301 caagacagca tcaaacctca ctgtctctgt cctggaagca gaaggagtct ttgaaaaagg
      361 ctcttttaag acacttactc ttccatcact acctctgaac agtgcagatg agatttatga
      421 gctacgtgta accggacgta cccaggatga gattttattc tctaatagta cccgcttatc
      481 atttgagacc aagagaatat ctgtcttcat tcaaacagac aaggccttat acaagccaaa
      541 gcaagaagtg aagtttcgca ttgttacact cttctcagat tttaagcctt acaaaacctc
      601 tttaaacatt ctcattaagg accccaaatc aaatttgatc caacagtggt tgtcacaaca
      661 aagtgatctt ggagtcattt ccaaaacttt tcagctatct tcccatccaa tacttggtga
      721 ctggtctatt caagttcaag tgaatgacca gacatattat caatcatttc aggtttcaga
      781 atatgtatta ccaaaatttg aagtgacttt gcagacacca ttatattgtt ctatgaattc
      841 taagcattta aatggtacca tcacggcaaa gtatacatat gggaagccag tgaaaggaga
      901 cgtaacgctt acatttttac ctttatcctt ttggggaaag aagaaaaata ttacaaaaac
      961 atttaagata aatggatctg caaacttctc ttttaatgat gaagagatga aaaatgtaat
     1021 ggattcttca aatggacttt ctgaatacct ggatctatct tcccctggac cagtagaaat
     1081 tttaaccaca gtgacagaat cagttacagg tatttcaaga aatgtaagca ctaatgtgtt
     1141 cttcaagcaa catgattaca tcattgagtt ttttgattat actactgtct tgaagccatc
     1201 tctcaacttc acagccactg tgaaggtaac tcgtgctgat ggcaaccaac tgactcttga
     1261 agaaagaaga aataatgtag tcataacagt gacacagaga aactatactg agtactggag
     1321 cggatctaac agtggaaatc agaaaatgga agctgttcag aaaataaatt atactgtccc
     1381 ccaaagtgga acttttaaga ttgaattccc aatcctggag gattccagtg agctacagtt
     1441 gaaggcctat ttccttggta gtaaaagtag catggcagtt catagtctgt ttaagtctcc
     1501 tagtaagaca tacatccaac taaaaacaag agatgaaaat ataaaggtgg gatcgccttt
     1561 tgagttggtg gttagtggca acaaacgatt gaaggagtta agctatatgg tagtatccag
     1621 gggacagttg gtggctgtag gaaaacaaaa ttcaacaatg ttctctttaa caccagaaaa
     1681 ttcttggact ccaaaagcct gtgtaattgt gtattatatt gaagatgatg gggaaattat
     1741 aagtgatgtt ctaaaaattc ctgttcagct tgtttttaaa aataagataa agctatattg
     1801 gagtaaagtg aaagctgaac catctgagaa agtctctctt aggatctctg tgacacagcc
     1861 tgactccata gttgggattg tagctgttga caaaagtgtg aatctgatga atgcctctaa
     1921 tgatattaca atggaaaatg tggtccatga gttggaactt tataacacag gatattattt
     1981 aggcatgttc atgaattctt ttgcagtctt tcaggaatgt ggactctggg tattgacaga
     2041 tgcaaacctc acgaaggatt atattgatgg tgtttatgac aatgcagaat atgctgagag
     2101 gtttatggag gaaaatgaag gacatattgt agatattcat gacttttctt tgggtagcag
     2161 tccacatgtc cgaaagcatt ttccagagac ttggatttgg ctagacacca acatgggtta
```

```
2221 caggatttac caagaatttg aagtaactgt acctgattct atcacttctt gggtggctac
2281 tggttttgtg atctctgagg acctgggtct tggactaaca actactccag tggagctcca
2341 agccttccaa ccatttttca ttttttttgaa tcttccctac tctgttatca gaggtgaaga
2401 atttgctttg gaaataacta tattcaatta tttgaaagat gccactgagg ttaaggtaat
2461 cattgagaaa agtgacaaat ttgatattct aatgacttca aatgaaataa atgccacagg
2521 ccaccagcag acccttctgg ttcccagtga ggatggggca actgttcttt ttcccatcag
2581 gccaacacat ctgggagaaa ttcctatcac agtcacagct ctttcaccca ctgcttctga
2641 tgctgtcacc cagatgattt tagtaaaggc tgaaggaata gaaaaatcat attcacaatc
2701 catcttatta gacttgactg acaataggct acagagtacc ctgaaaactt tgagtttctc
2761 atttcctcct aatacagtga ctggcagtga aagagttcag atcactgcaa ttggagatgt
2821 tcttggtcct tccatcaatg gcttagcctc attgattcgg atgccttatg gctgtggtga
2881 acagaacatg ataaattttg ctccaaatat ttacattttg gattatctga ctaaaaagaa
2941 acaactgaca gataatttga aagaaaaagc tctttcattt atgaggcaag gttaccagag
3001 agaacttctc tatcagaggg aagatggctc tttcagtgct tttgggaatt atgacccttc
3061 tgggagcact tggttgtcag cttttgtttt aagatgtttc cttgaagccg atccttacat
3121 agatattgat cagaatgtgt tacacagaac atacacttgg cttaaaggac atcagaaatc
3181 caacggtgaa ttttgggatc caggaagagt gattcatagt gagcttcaag gtggcaataa
3241 aagtccagta acacttacag cctatattgt aacttctctc ctgggatata gaaagtatca
3301 gcctaacatt gatgtgcaag agtctatcca ttttttggag tctgaattca gtagaggaat
3361 ttcagacaat tatactctag cccttataac ttatgcattg tcatcagtgg ggagtcctaa
3421 agcgaaggaa gctttgaata tgctgacttg gagagcagaa caagaaggtg gcatgcaatt
3481 ctgggtgtca tcagagtcca aactttctga ctcctggcag ccacgctccc tggatattga
3541 agttgcagcc tatgcactgc tctcacactt cttacaattt cagacttctg agggaatccc
3601 aattatgagg tggctaagca ggcaaagaaa tagcttgggt ggttttgcat ctactcagga
3661 taccactgtg gctttaaagg ctctgtctga atttgcagcc ctaatgaata cagaaaggac
3721 aaatatccaa gtgaccgtga cggggcctag ctcaccaagt cctgtaaagt ttctgattga
3781 cacacacaac cgcttactcc ttcagacagc agagcttgct gtggtacagc caatggcagt
3841 taatatttcc gcaaatggtt ttggatttgc tatttgtcag ctcaatgttg tatataatgt
3901 gaaggcttct gggtcttcta gaagacgaag atctatccaa aatcaagaag cctttgattt
3961 agatgttgct gtaaaagaaa ataaagatga tctcaatcat gtggatttga atgtgtgtac
4021 aagcttttcg ggcccgggta ggagtggcat ggctcttatg gaagttaacc tattaagtgg
4081 ctttatggtg ccttcagaag caatttctct gagcgagaca gtgaagaaag tggaatatga
4141 tcatggaaaa ctcaacctct atttagattc tgtaaatgaa acccagtttt gtgttaatat
4201 tcctgctgtg agaaacttta aagtttcaaa tacccaagat gcttcagtgt ccatagtgga
4261 ttactatgag ccaaggagac aggcggtgag aagttacaac tctgaagtga agctgtcctc
4321 ctgtgacctt tgcagtgatg tccagggctg ccgtccttgt gaggatggag cttcaggctc
4381 ccatcatcac tcttcagtca tttttatttt ctgtttcaag cttctgtact ttatggaact
4441 ttggctgtga tttatttta aaggactctg tgtaacacta acatttccag tagtcacatg
4501 tgattgtttt gttttcgtag aagaatactg cttctatttt gaaaaaagag ttttttttct
4561 ttctatgggg ttgcagggat ggtgtacaac aggtcctagc atgtatagct gcatagattt
4621 cttcacctga tctttgtgtg gaagatcaga atgaatgcag ttgtgtgtct atattttccc
4681 ctcacaaaat cttttagaat ttttttggag gtgtttgttt tctccagaat aaaggtatta
4741 ctttagaaat aggtattctc ctcattttgt gaaagaaatg aacctagatt cttaagcatt
4801 attacacatc catgtttgct taaagatgga tttccctggg aatgggagaa aacagccagc
4861 aggaggagct tcatctgttc ccttcccacc tccaacctag ccctactgcc caccccaccc
4921 caacccaccc catgcccagt ggtctcagta gatacttctt aactggaaat tctttctttt
4981 cagaatctag gtggtgaatt ttttttaagt ggcacggtct ttttctgctt gaaatctgat
5041 cacacccccc agccattgcc ctccctctct ttttcctctg tagagaaatg tgaggggcag
5101 tacatttact gtgcttttca caccatctca gaggttgagg agcatactga aaattgccct
5161 ggggggtgct gggtgtgctg tctccttccc acatcctcag ccccacacca gctctatttc
5221 aggggtgaga gtcagagagc actgcaatat gtgcttcatg ggatttcgat tcgaagatcc
5281 tagaccaggg agacactgtg agccagggat acaacaaaat actaggtaag tcactgcaga
5341 ccgacctccc tgcagtttgg gaaagaagct gggtttgtgg agaatcagag catcttgaca
5401 tgactgctga cctaaagatc cctggcattg gccagggatc ctgtggaacc tcttctagtt
5461 caggggtgtg agcattagac tgccagttgt ctagtgacat ctgatgcttg ctgtgaactt
5521 ttaagatccc cgaatcctga gcacctcaat ctttaattgc cctgtattcc gaagggtaat
5581 ataatttatc tggatggaaa ttttaaagat gaatcccccc ttttttcttt cttctctctt
5641 ttctttcctt ctccctttct tctttgcctt ctaaatatac tgaaatgatt tagatatgtg
5701 tcaacaatta atgatctttt attcaatcta agaaatggtt tagtttttct ctttagctct
5761 atggcatttc actcaagtgg acaggggaaa aagtaattgc catgggctcc aaagaatttg
```

```
     5821 ctttatgttt ttagctattt aaaaataaat ccatcaaaaa taaagtatgc aaatgtatct
     5881 ttt
//
```


May 12 2004 07:05:19

>AY149920 ACCESSION:AY149920 NID: gi 37359235 gb AY149920.1 Homo sapiens activated T-cell marker CD109 (CD109) mRNA, complete cds
Length = 4338

Score = 2845 bits (7294), Expect = 0.0
Identities = 1423/1445 (98%), Positives = 1425/1445 (98%), Gaps = 17/1445 (1%)
Frame = +1

```
Query: 1    MQGPPLLTAAHLLCVCTAALAVAPGPRFLVTAPGIIRPGGNVTIGVELLEHCPSQVTVKA 60
            MQGPPLLTAAHLLCVCTAALAVAPGPRFLVTAPGIIRPGGNVTIGVELLEHCPSQVTVKA
Sbjct: 1    MQGPPLLTAAHLLCVCTAALAVAPGPRFLVTAPGIIRPGGNVTIGVELLEHCPSQVTVKA 180

Query: 61   ELLKTASNLTVSVLEAEGVFEKGSFKTLTLPSLPLNSADEIYELRVTGRTQDEILFSNST 120
            ELLKTASNLTVSVLEAEGVFEKGSFKTLTLPSLPLNSADEIYELRVTGRTQDEILFSNST
Sbjct: 181  ELLKTASNLTVSVLEAEGVFEKGSFKTLTLPSLPLNSADEIYELRVTGRTQDEILFSNST 360

Query: 121  RLSFETKRISVFIQTDKALYKPKQEVKFRIVTLFSDFKPYKTSLNILIKDPKSNLIQQWL 180
            RLSFETKRISVFIQTDKALYKPKQEVKFRIVTLFSDFKPYKTSLNILIKDPKSNLIQQWL
Sbjct: 361  RLSFETKRISVFIQTDKALYKPKQEVKFRIVTLFSDFKPYKTSLNILIKDPKSNLIQQWL 540

Query: 181  SQQSDLGVISKTFQLSSHPILGDWSIQVQVNDQTYYQSFQVSEYVLPKFEVTLQTPLYCS 240
            SQQSDLGVISKTFQLSSHPILGDWSIQVQVNDQTYYQSFQVSEYVLPKFEVTLQTPLYCS
Sbjct: 541  SQQSDLGVISKTFQLSSHPILGDWSIQVQVNDQTYYQSFQVSEYVLPKFEVTLQTPLYCS 720

Query: 241  MNSKHLNGTITAKYTYGKPVKGDVTLTFLPLSFWGKKKNITKTFKINGSANFSFNDEEMK 300
            MNSKHLNGTITAKYTYGKPVKGDVTLTFLPLSFWGKKKNITKTFKINGSANFSFNDEEMK
Sbjct: 721  MNSKHLNGTITAKYTYGKPVKGDVTLTFLPLSFWGKKKNITKTFKINGSANFSFNDEEMK 900

Query: 301  NVMDSSNGLSEYLDLSSPGPVEILTTVTESVTGISRNVSTNVFFKQHDYIIEFFDYTTVL 360
            NVMDSSNGLSEYLDLSSPGPVEILTTVTESVTGISRNVSTNVFFKQHDYIIEFFDYTTVL
Sbjct: 901  NVMDSSNGLSEYLDLSSPGPVEILTTVTESVTGISRNVSTNVFFKQHDYIIEFFDYTTVL 1080

Query: 361  KPSLNFTATVKVTRADGNQLTLEERRNNVVITVTQRNYTEYWSGSNSGNQKMEAVQKINY 420
            KPSLNFTATVKVTRADGNQLTLEERRNNVVITVTQRNYTEYWSGSNSGNQKMEAVQKINY
Sbjct: 1081 KPSLNFTATVKVTRADGNQLTLEERRNNVVITVTQRNYTEYWSGSNSGNQKMEAVQKINY 1260

Query: 421  TVPQSGTFKIEFPILEDSSELQLKAYFLGSKSSMAVHSLFKSPSKTYIQLKTRDENIKVG 480
            TVPQSGTFKIEFPILEDSSELQLKAYFLGSKSSMAVHSLFKSPSKTYIQLKTRDENIKVG
Sbjct: 1261 TVPQSGTFKIEFPILEDSSELQLKAYFLGSKSSMAVHSLFKSPSKTYIQLKTRDENIKVG 1440

Query: 481  SPFELVVSGNKRLKELSYMVVSRGQLVAVGKQNSTMFSLTPENSWTPKACVIVYYIEDDG 540
            SPFELVVSGNKRLKELSYMVVSRGQLVAVGKQNSTMFSLTPENSWTPKACVIVYYIEDDG
Sbjct: 1441 SPFELVVSGNKRLKELSYMVVSRGQLVAVGKQNSTMFSLTPENSWTPKACVIVYYIEDDG 1620

Query: 541  EIISDVLKIPVQLVFKNKIKLYWSKVKAEPSEKVSLRISVTQPDSIVGIVAVDKSVNLMN 600
            EIISDVLKIPVQLVFKNKIKLYWSKVKAEPSEKVSLRISVTQPDSIVGIVAVDKSVNLMN
Sbjct: 1621 EIISDVLKIPVQLVFKNKIKLYWSKVKAEPSEKVSLRISVTQPDSIVGIVAVDKSVNLMN 1800

Query: 601  ASNDITMENVVHELELYNTGYYLGMFMNSFAVFQECGLWVLTDANLTKDYIDGVYDNAEY 660
            ASNDITMENVVHELELYNTGYYLGMF+NSFAVFQECGLWVLTDANLTKDYIDGVYDNAEY
Sbjct: 1801 ASNDITMENVVHELELYNTGYYLGMFINSFAVFQECGLWVLTDANLTKDYIDGVYDNAEY 1980

Query: 661  AERFMEENEGHIVDIHDFSLGSSPHVRKHFPETWIWLDTNMGYRIYQEFEVTVPDSITSW 720
            AERFMEENEGHIVDIHDFSLGSSPHVRKHFPETWIWLDTNMG RIYQEFEVTVPDSITSW
Sbjct: 1981 AERFMEENEGHIVDIHDFSLGSSPHVRKHFPETWIWLDTNMGSRIYQEFEVTVPDSITSW 2160
```

```
Query: 721  VATGFVISEDLGLGLTTTPVELQAFQPFFIFLNLPYSVIRGEEFALEITIFNYLKDATEV 780
            VATGFVISEDLGLGLTTTPVELQAFQPFFIFLNLPYSVIRGEEFALEITIFNYLKDATEV
Sbjct: 2161 VATGFVISEDLGLGLTTTPVELQAFQPFFIFLNLPYSVIRGEEFALEITIFNYLKDATEV 2340

Query: 781  KVIIEKSDKFDILMTSSEINATGHQQTLLVPSEDGATVLFPIRPTHLGEIPITVTALSPT 840
            KVIIEKSDKFDILMTSSEINAT HQQTLLVPSEDGATVLFPIRPTHLGEIPITVTALSPT
Sbjct: 2341 KVIIEKSDKFDILMTSSEINATSHQQTLLVPSEDGATVLFPIRPTHLGEIPITVTALSPT 2520

Query: 841  ASDAVTQMILVKAEGIEKSYSQSILLDLTDNRLQSTLKTLSFSFPPNTVTGSERVQITAI 900
            ASDA+TQMILVKAEGIEKSYSQSILLDLTDNRLQSTLKTLSFSFPPNTVTGSERVQITAI
Sbjct: 2521 ASDAITQMILVKAEGIEKSYSQSILLDLTDNRLQSTLKTLSFSFPPNTVTGSERVQITAI 2700

Query: 901  GDVLGPSINGLASLIRMPYGCGEQNMINFAPNIYILDYLTKKKQLTDNLKEKALSFMRQG 960
            GDVLGPSINGLASLIRMPYGCGEQNMINFAPNIYILDYLTKKKQLTDNLKEKALSFMRQG
Sbjct: 2701 GDVLGPSINGLASLIRMPYGCGEQNMINFAPNIYILDYLTKKKQLTDNLKEKALSFMRQG 2880

Query: 961  YQRELLYQREDGSFSAFGNYDPSGSTWLSAFVLRCFLEADPYIDIDQNVLHRTYTWLKGH 1020
            YQRELLYQREDGSFSAFGNYDPSGSTWLSAFVLRCFLEADPYIDIDQNVLHRTYTWLKGH
Sbjct: 2881 YQRELLYQREDGSFSAFGNYDPSGSTWLSAFVLRCFLEADPYIDIDQNVLHRTYTWLKGH 3060

Query: 1021 QKSNGEFWDPGRVIHSELQGGNKSPVTLTAYIVTSLLGYRKYQPNIDVQESIHFLESEFS 1080
            QKSNGEFWDPGRVIHSELQGGNKSPVTLTAYIVTSLLGYRKYQPNIDVQESIHFLESEFS
Sbjct: 3061 QKSNGEFWDPGRVIHSELQGGNKSPVTLTAYIVTSLLGYRKYQPNIDVQESIHFLESEFS 3240

Query: 1081 RGISDNYTLALITYALSSVGSPKAKEALNMLTWRAEQEGGMQFWVSSESKLSDSWQPRSL 1140
            RGISDNYTLALITYALSSVGSPKAKEALNMLTWRAEQEGGMQFWVSSESKLSDSWQPRSL
Sbjct: 3241 RGISDNYTLALITYALSSVGSPKAKEALNMLTWRAEQEGGMQFWVSSESKLSDSWQPRSL 3420

Query: 1141 DIEVAAYALLSHFLQFQTSEGIPIMRWLSRQRNSLGGFASTQDTTVALKALSEFAALMNT 1200
            DIEVAAYALLSHFLQFQTSEGIPIMRWLSRQRNSLGGFASTQDTTVALKALSEFAALMNT
Sbjct: 3421 DIEVAAYALLSHFLQFQTSEGIPIMRWLSRQRNSLGGFASTQDTTVALKALSEFAALMNT 3600

Query: 1201 ERTNIQVTVTGPSSPSP----------------LAVVQPMAVNISANGFGFAICQLNVV 1243
            ERTNIQVTVTGPSSPSP                  LAVVQP AVNISANGFGFAICQLNVV
Sbjct: 3601 ERTNIQVTVTGPSSPSPVKFLIDTHNRLLLQTAELAVVQPTAVNISANGFGFAICQLNVV 3780

Query: 1244 YNVKASGSSRRRRSIQNQEAFDLDVAVKENKDDLNHVDLNVCTSFSGPGRSGMALMEVNL 1303
            YNVKASGSSRRRRSIQNQEAFDLDVAVKENKDDLNHVDLNVCTSFSGPGRSGMALMEVNL
Sbjct: 3781 YNVKASGSSRRRRSIQNQEAFDLDVAVKENKDDLNHVDLNVCTSFSGPGRSGMALMEVNL 3960

Query: 1304 LSGFMVPSEAISLSETVKKVEYDHGKLNLYLDSVNETQFCVNIPAVRNFKVSNTQDASVS 1363
            LSGFMVPSEAISLSETVKKVEYDHGKLNLYLDSVNETQFCVNIPAVRNFKVSNTQDASVS
Sbjct: 3961 LSGFMVPSEAISLSETVKKVEYDHGKLNLYLDSVNETQFCVNIPAVRNFKVSNTQDASVS 4140

Query: 1364 IVDYYEPRRQAVRSYNSEVKLSSCDLCSDVQGCRPCEDGASGSHHHSSVIFIFCFKLLYF 1423
            IVDYYEPRRQAVRSYNSEVKLSSCDLCSDVQGCRPCEDGASGSHHHSSVIFIFCFKLLYF
Sbjct: 4141 IVDYYEPRRQAVRSYNSEVKLSSCDLCSDVQGCRPCEDGASGSHHHSSVIFIFCFKLLYF 4320

Query: 1424 MELWL 1428
            MELWL
Sbjct: 4321 MELWL 4335
```


```
LOCUS       AY149920                4338 bp    mRNA    linear   PRI 06-APR-2004
DEFINITION  Homo sapiens activated T-cell marker CD109 (CD109) mRNA, complete
            cds.
ACCESSION   AY149920
VERSION     AY149920.1  GI:37359235
KEYWORDS    .
SOURCE      Homo sapiens (human)
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.
REFERENCE   1  (bases 1 to 4338)
  AUTHORS   Solomon,K.R., Sharma,P., Chan,M., Morrison,P.T. and Finberg,R.W.
  TITLE     CD109 represents a novel branch of the
            alpha2-macroglobulin/complement gene family
  JOURNAL   Gene 327 (2), 171-183 (2004)
   PUBMED   14980714
REFERENCE   2  (bases 1 to 4338)
  AUTHORS   Solomon,K., Sharma,P., Morrison,P. and Finberg,R.W.
  TITLE     Direct Submission
  JOURNAL   Submitted (12-SEP-2002) Medicine, UMass Medical School, 364
            Plantation Street, Worcester, MA 01605, USA
FEATURES             Location/Qualifiers
     source          1..4338
                     /organism="Homo sapiens"
                     /mol_type="mRNA"
                     /db_xref="taxon:9606"
                     /chromosome="6"
                     /map="6q"
                     /cell_line="U373"
     gene            1..4338
                     /gene="CD109"
     CDS             1..4338
                     /gene="CD109"
                     /note="AMCOM; alpha2-macroglobulin/complement;
                     GPI-anchored to membrane; hematopoetic stem cell marker;
                     inducible by PMA, LPS and phytohemaglutinin"
                     /codon_start=1
                     /product="activated T-cell marker CD109"
                     /protein_id="AAN78483.1"
                     /db_xref="GI:37359236"
                     /translation="MQGPPLLTAAHLLCVCTAALAVAPGPRFLVTAPGIIRPGGNVTI
                     GVELLEHCPSQVTVKAELLKTASNLTVSVLEAEGVFEKGSFKTLTLPSLPLNSADEIY
                     ELRVTGRTQDEILFSNSTRLSFETKRISVFIQTDKALYKPKQEVKFRIVTLFSDFKPY
                     KTSLNILIKDPKSNLIQQWLSQQSDLGVISKTFQLSSHPILGDWSIQVQVNDQTYYQS
                     FQVSEYVLPKFEVTLQTPLYCSMNSKHLNGTITAKYTYGKPVKGDVTLTFLPLSFWGK
                     KKNITKTFKINGSANFSFNDEEMKNVMDSSNGLSEYLDLSSPGPVEILTTVTESVTGI
                     SRNVSTNVFFKQHDYIIEFFDYTTVLKPSLNFTATVKVTRADGNQLTLEERRNNVVIT
                     VTQRNYTEYWSGSNSGNQKMEAVQKINYTVPQSGTFKIEFPILEDSSELQLKAYFLGS
                     KSSMAVHSLFKSPSKTYIQLKTRDENIKVGSPFELVVSGNKRLKELSYMVVSRGQLVA
```

VGKQNSTMFSLTPENSWTPKACVIVYYIEDDGEIISDVLKIPVQLVFKNKIKLYWSKV
KAEPSEKVSLRISVTQPDSIVGIVAVDKSVNLMNASNDITMENVVHELELYNTGYYLG
MFINSFAVFQECGLWVLTDANLTKDYIDGVYDNAEYAERFMEENEGHIVDIHDFSLGS
SPHVRKHFPETWIWLDTNMGSRIYQEFEVTVPDSITSWVATGFVISEDLGLGLTTTPV
ELQAFQPFFIFLNLPYSVIRGEEFALEITIFNYLKDATEVKVIIEKSDKFDILMTSSE
INATSHQQTLLVPSEDGATVLFPIRPTHLGEIPITVTALSPTASDAITQMILVKAEGI
EKSYSQSILLDLTDNRLQSTLKTLSFSFPPNTVTGSERVQITAIGDVLGPSINGLASL
IRMPYGCGEQNMINFAPNIYILDYLTKKKQLTDNLKEKALSFMRQGYQRELLYQREDG
SFSAFGNYDPSGSTWLSAFVLRCFLEADPYIDIDQNVLHRTYTWLKGHQKSNGEFWDP
GRVIHSELQGGNKSPVTLTAYIVTSLLGYRKYQPNIDVQESIHFLESEFSRGISDNYT
LALITYALSSVGSPKAKEALNMLTWRAEQEGGMQFWVSSESKLSDSWQPRSLDIEVAA
YALLSHFLQFQTSEGIPIMRWLSRQRNSLGGFASTQDTTVALKALSEFAALMNTERTN
IQVTVTGPSSPSPVKFLIDTHNRLLLQTAELAVVQPTAVNISANGFGFAICQLNVVYN
VKASGSSRRRRSIQNQEAFDLDVAVKENKDDLNHVDLNVCTSFSGPGRSGMALMEVNL
LSGFMVPSEAISLSETVKKVEYDHGKLNLYLDSVNETQFCVNIPAVRNFKVSNTQDAS
VSIVDYYEPRRQAVRSYNSEVKLSSCDLCSDVQGCRPCEDGASGSHHHSSVIFIFCFK
LLYFMELWL"

ORIGIN

```
        1 atgcagggcc caccgctcct gaccgccgcc cacctcctct gcgtgtgcac cgccgcgctg
       61 gccgtggctc ccgggcctcg gtttctggtg acagccccag ggatcatcag gcccggagga
      121 aatgtgacta ttggggtgga gcttctggaa cactgccctt cacaggtgac tgtgaaggcg
      181 gagctgctca agacagcatc aaacctcact gtctctgtcc tggaagcaga aggagtcttt
      241 gaaaaaggct cttttaagac acttactctt ccatcactac ctctgaacag tgcagatgag
      301 atttatgagc tacgtgtaac cggacgtacc caggatgaga ttttattctc taatagtacc
      361 cgcttatcat ttgagaccaa gagaatatct gtcttcattc aaacagacaa ggccttatac
      421 aagccaaagc aagaagtgaa gtttcgcatt gttacactct tctcagattt taagccttac
      481 aaaacctctt taaacattct cattaaggac cccaaatcaa atttgatcca acagtggttg
      541 tcacaacaaa gtgatcttgg agtcatttcc aaaacttttc agctatcttc ccatccaata
      601 cttggtgact ggtctattca agttcaagtg aatgaccaga catattatca atcatttcag
      661 gtttcagaat atgtattacc aaaatttgaa gtgactttgc agacaccatt atattgttct
      721 atgaattcta agcatttaaa tggtaccatc acggcaaagt atacatatgg gaagccagtg
      781 aaaggagacg taacgcttac atttttacct ttatcctttt ggggaaagaa gaaaaatatt
      841 acaaaaacat ttaagataaa tggatctgca aacttctctt ttaatgatga agagatgaaa
      901 aatgtaatgg attcttcaaa tggactttct gaatacctgg atctatcttc ccctggacca
      961 gtagaaattt taaccacagt gacagaatca gttacaggta tttcaagaaa tgtaagcact
     1021 aatgtgttct tcaagcaaca tgattacatc attgagtttt ttgattatac tactgtcttg
     1081 aagccatctc tcaacttcac agccactgtg aaggtaactc gtgctgatgg caaccaactg
     1141 actcttgaag aaagaagaaa taatgtagtc ataacagtga cacagagaaa ctatactgag
     1201 tactggagcg gatctaacag tggaaatcag aaaatggaag ctgttcagaa aataaattat
     1261 actgtccccc aaagtggaac ttttaagatt gaattcccaa tcctggagga ttccagtgag
     1321 ctacagttga aggcctattt ccttggtagt aaaagtagca tggcagttca tagtctgttt
     1381 aagtctccta gtaagacata catccaacta aaaacaagag atgaaaatat aaaggtggga
     1441 tcgccttttg agttggtggt tagtggcaac aaacgattga aggagttaag ctatatggta
     1501 gtatccaggg gacagttggt ggctgtagga aaacaaaatt caacaatgtt ctctttaaca
     1561 ccagaaaatt cttggactcc aaaagcctgt gtaattgtgt attatattga agatgatggg
     1621 gaaattataa gtgatgttct aaaaattcct gttcagcttg tttttaaaaa taagataaag
     1681 ctatattgga gtaaagtgaa agctgaacca tctgagaaag tctctcttag gatctctgtg
     1741 acacagcctg actccatagt tgggattgta gctgttgaca aaagtgtgaa tctgatgaat
     1801 gcctctaatg atattacaat ggaaaatgtg gtccatgagt tggaacttta taacacagga
     1861 tattatttag gcatgttcat aaattctttt gcagtctttc aggaatgtgg actctgggta
     1921 ttgacagatg caaacctcac gaaggattat attgatggtg tttatgacaa tgcagaatat
     1981 gctgagaggt ttatggagga aaatgaagga catattgtag atattcatga cttttctttg
     2041 ggtagcagtc cacatgtccg aaagcatttt ccagagactt ggatttggct agacaccaac
     2101 atgggttcca ggatttacca agaatttgaa gtaactgtac ctgattctat cacttcttgg
     2161 gtggctactg gttttgtgat ctctgaggac ctgggtcttg gactaacaac tactccagtg
     2221 gagctccaag ccttccaacc atttttcatt tttttgaatc ttccctactc tgttatcaga
     2281 ggtgaagaat ttgctttgga aataactata ttcaattatt tgaaagatgc cactgaggtt
     2341 aaggtaatca ttgagaaaag tgacaaattt gatattctaa tgacttcaag tgaaataaat
     2401 gccacaagcc accagcagac ccttctggtt cccagtgagg atggggcaac tgttcttttt
     2461 cccatcaggc caacacatct gggagaaatt cctatcacag tcacagctct ttcacccact
```

```
     2521 gcttctgatg ctatcaccca gatgatttta gtaaaggctg aaggaataga aaaatcatat
     2581 tcacaatcca tcttattaga cttgactgac aataggctac agagtaccct gaaaactttg
     2641 agtttctcat ttcctcctaa tacagtgact ggcagtgaaa gagttcagat cactgcaatt
     2701 ggagatgttc ttggtccttc catcaatggc ttagcctcat tgattcggat gccttatggc
     2761 tgtggtgaac agaacatgat aaattttgct ccaaatattt acattttgga ttatctgact
     2821 aaaaagaaac aactgacaga taatttgaaa gaaaaagctc tttcatttat gaggcaaggt
     2881 taccagagag aacttctcta tcagagggaa gatggctctt tcagtgcttt tgggaattat
     2941 gacccttctg ggagcacttg gttgtcagct tttgttttaa gatgtttcct tgaagccgat
     3001 ccttacatag atattgatca gaatgtgtta cacagaacat acacttggct taaaggacat
     3061 cagaaatcca acggtgaatt ttgggatcca ggaagagtga ttcatagtga gcttcaaggt
     3121 ggcaataaaa gtccagtaac acttacagcc tatattgtaa cttctctcct gggatataga
     3181 aagtatcagc ctaacattga tgtgcaagag tctatccatt ttttggagtc tgaattcagt
     3241 agaggaattt cagacaatta tactctagcc cttataactt atgcattgtc atcagtgggg
     3301 agtcctaaag cgaaggaagc tttgaatatg ctgacttgga gagcagaaca agaaggtggc
     3361 atgcaattct gggtgtcatc agagtccaaa ctttctgact cctggcagcc acgctccctg
     3421 gatattgaag ttgcagccta tgcactgctc tcacacttct tacaatttca gacttctgag
     3481 ggaatcccaa ttatgaggtg gctaagcagg caaagaaata gcttgggtgg ttttgcatct
     3541 actcaggata ccactgtggc tttaaaggct ctgtctgaat ttgcagccct aatgaataca
     3601 gaaaggacaa atatccaagt gaccgtgacg gggcctagct caccaagtcc tgtaaagttt
     3661 ctgattgaca cacacaaccg cttactcctt cagacagcag agcttgctgt ggtacagcca
     3721 acggcagtta atatttccgc aaatggtttt ggatttgcta tttgtcagct caatgttgta
     3781 tataatgtga aggcttctgg gtcttctaga agacgaagat ctatccaaaa tcaagaagcc
     3841 tttgatttag atgttgctgt aaaagaaaat aaagatgatc tcaatcatgt ggatttgaat
     3901 gtgtgtacaa gcttttcggg cccgggtagg agtggcatgg ctcttatgga agttaaccta
     3961 ttaagtggct ttatggtgcc ttcagaagca atttctctga gcgagacagt gaagaaagtg
     4021 gaatatgatc atggaaaact caacctctat ttagattctg taaatgaaac ccagttttgt
     4081 gttaatattc ctgctgtgag aaactttaaa gtttcaaata cccaagatgc ttcagtgtct
     4141 atagtggatt actatgagcc aaggagacag gcggtgagaa gttacaactc tgaagtgaag
     4201 ctgtcctcct gtgacctttg cagtgatgtc cagggctgcc gtccttgtga ggatggagct
     4261 tcaggctccc atcatcactc ttcagtcatt tttattttct gtttcaagct tctgtacttt
     4321 atggaacttt ggctgtga
//
```

Cell surface antigen CD109 is a novel member of the alpha(2) macroglobulin/C3, C4, C5 family of thioester-containing proteins.

Lin M, Sutherland DR, Horsfall W, Totty N, Yeo E, Nayar R, Wu XF, Schuh AC.

Department of Medical Biophysics, University of Toronto, ON, Canada.

Cell surface antigen CD109 is a glycosylphosphatidylinositol (GPI)-linked glycoprotein of approximately 170 kd found on a subset of hematopoietic stem and progenitor cells and on activated platelets and T cells. Although it has been suggested that T-cell CD109 may play a role in antibody-inducing T-helper function and it is known that platelet CD109 carries the Gov alloantigen system, the role of CD109 in hematopoietic cells remains largely unknown. As a first step toward elucidating the function of CD109, we have isolated and characterized a human CD109 cDNA from KG1a and endothelial cells. The isolated cDNA comprises a 4335 bp open-reading frame encoding a 1445 amino acid (aa) protein of approximately 162 kd that contains a 21 aa N-terminal leader peptide, 17 potential N-linked glycosylation sites, and a C-terminal GPI anchor cleavage-addition site. We report that CD109 is a novel member of the alpha 2 macroglobulin (alpha 2M)/C3, C4, C5 family of thioester-containing proteins, and we demonstrate that native CD109 does indeed contain an intact thioester. Analysis of the CD109 aa sequence suggests that CD109 is likely activated by proteolytic cleavage and thereby becomes capable of thioester-mediated covalent binding to adjacent molecules or cells. In addition, the predicted chemical reactivity of the activated CD109 thioester is complement-like rather than resembling that of alpha 2M proteins. Thus, not only is CD109 potentially capable of covalent binding to carbohydrate and protein targets, but the t(1/2) of its activated thioester is likely extremely short, indicating that CD109 action is highly restricted spatially to the site of its activation.

MeSH Terms:

- Amino Acid Sequence
- Antigens, CD*/chemistry
- Antigens, CD*/genetics
- Antigens, CD*/metabolism
- Base Sequence

- Cysteine
- DNA, Complementary/analysis
- DNA, Complementary/genetics
- DNA, Complementary/isolation & purification
- Glutamine
- Glycosylphosphatidylinositols/chemistry
- Hematopoietic Stem Cells/chemistry
- Hematopoietic Stem Cells/immunology
- Human
- Molecular Sequence Data
- Sequence Alignment
- Sequence Analysis, DNA
- Sulfides/chemistry
- Support, Non-U.S. Gov't
- Tumor Cells, Cultured
- Variation (Genetics)
- alpha-Macroglobulins/chemistry*
- alpha-Macroglobulins/genetics
- alpha-Macroglobulins/metabolism

Substances:

- Antigens, CD
- CD109 antigen, human
- DNA, Complementary
- Glycosylphosphatidylinositols
- Sulfides
- alpha-Macroglobulins
- Cysteine
- Glutamine

Secondary Source ID:

- GENBANK/AF410459

PMID: 11861284 [PubMed - indexed for MEDLINE]

1: Gene. 2004 Mar 3;327(2):171-83.

CD109 represents a novel branch of the alpha2-macroglobulin/complement gene family.

Solomon KR, Sharma P, Chan M, Morrison PT, Finberg RW.

Department of Orthopaedic Surgery, Children's Hospital, Boston, MA 02115, USA.

We report here the genomic organization and phylogenic relationships of CD109, a member of the the alpha2-macroglobulin/complement (AMCOM) gene family. CD109 is a GPI-linked glycoprotein expressed on endothelial cells, platelets, activated T-cells, and a wide variety of tumors. We cloned full-length CD109 cDNA from the mammalian U373 cell line by RT-PCR and performed analysis of its corresponding genomic sequence. The CD109 cDNA spans 128 kb of chromosome 6q with its 33 exons constituting approximately 3.3% of the total CD109 genomic sequence. Sequence analysis revealed that CD109 contains specific motifs in its N-terminus, that are highly conserved in all AMCOM members. CD109 also shares motifs with certain other AMCOM members including: (1) a thioester 'GCGEQ" motif, (2) a furin site of four positively charged amino acids, and (3) a double tyrosine near the C-terminus. Based on a phylogenic analysis of human CD109 with other human homologs as well as orthologs from other mammalian species, C. elegans (ZK337.1) and E. coli homologs, we propose CD109 represents a novel and independent branch of the alpha2-macroglobulin/complement gene family (AMCOM) and may be its oldest member.

MeSH Terms:

- Amino Acid Sequence
- Animals
- Antigens, CD/chemistry
- Antigens, CD/genetics*
- Antigens, CD/metabolism
- CHO Cells
- Cell Line, Tumor
- Chromosome Mapping
- Cloning, Molecular
- Complement/genetics*
- DNA, Complementary/chemistry

- DNA, Complementary/genetics
- Evolution, Molecular
- Exons
- Genes, Structural/genetics
- Hamsters
- Human
- Introns
- Molecular Sequence Data
- Multigene Family/genetics
- Phosphatidylinositol Diacylglycerol-Lyase/metabolism
- Phylogeny
- Reverse Transcriptase Polymerase Chain Reaction
- Sequence Alignment
- Sequence Analysis, DNA
- Sequence Analysis, Protein
- Sequence Homology, Amino Acid
- Support, Non-U.S. Gov't
- Support, U.S. Gov't, P.H.S.
- alpha-Macroglobulins/genetics*

Substances:

- Antigens, CD
- CD109 antigen, human
- DNA, Complementary
- alpha-Macroglobulins
- Complement
- Phosphatidylinositol Diacylglycerol-Lyase

Secondary Source ID:

- PIR/AY149920

Grant Support:

- R01 GM63244/GM/NIGMS

PMID: 14980714 [PubMed - indexed for MEDLINE]

App Serial # 10/020,095 Exhibit D
Walke et al. LEX-0282-USA
Novel Human Alpha Macroglobulin Family Proteins and Polynucleotides Encoding the Same

1: J Lab Clin Med. 1998 Aug;132(2):142-8.

ABH antigens on human platelets: expression on the glycosyl phosphatidylinositol-anchored protein CD109.

Kelton JG, Smith JW, Horsewood P, Warner MN, Warkentin TE, Finberg RW, Hayward CP.

Department of Medicine, McMaster University, Hamilton, Ontario, Canada.

Platelets express alloantigens that are platelet specific (eg, the HPA antigens) and alloantigens that are shared with other blood cells (eg, the ABH antigens). The blood group A and B determinants are expressed on glycolipids and on some intrinsic platelet membrane glycoproteins. This report characterizes multiple platelet proteins reacting with blood group antibodies in serum samples from mothers of children born with neonatal alloimmune thrombocytopenia. ABH antigens on additional platelet proteins are identified, including the glycosyl phosphatidylinositol-anchored protein CD109. The proteins that carry ABH antigens were identified by using monoclonal antibodies to glycoproteins Ib, IIb/IIIa, Ia/IIa, CD31, and CD109 and immunoprecipitation/immunoblotting techniques with monoclonal antibodies to A and B antigens. The maternal serum samples and anti-A and anti-B monoclonal antibodies immunoprecipitated identical radiolabeled platelet proteins including proteins at 220 and 175 kd and proteins with mobilities corresponding to glycoproteins Ib, IIb/IIIa, IV, and V. Treatment of platelets with phosphatidylinositol-specific phospholipase C released into the supernatant a 175-kd protein that expressed the blood group determinants. This protein comigrated with the glycosyl phosphatidylinositol-anchored protein CD109. When platelet proteins were purified by immunoprecipitation with monoclonal antibodies and then tested by immunoblotting, anti-A reacted with the glycosyl phosphatidylinositol-anchored protein CD109 and to glycoproteins Ib, IIb, IIa, IIIa, and CD31 (PECAM). These results indicate that structures for modification by glycosyltransferases exist on platelet CD109, which also expresses the Gov alloantigen system. This study indicates that certain platelet proteins express both platelet-specific and blood group antigens that may contribute to platelet transfusion refractoriness and to neonatal alloimmune thrombocytopenia.

PMID: 9708575 [PubMed - indexed for MEDLINE]

Display Abstract Show: 20 Sort Send to Text

Write to the Help Desk
NCBI | NLM | NIH
Department of Health & Human Services
Freedom of Information Act | Disclaimer

May 12 2004 06:43:50

1: Br J Haematol. 2000 Sep;110(3):735-42.

Detection of Gov system antibodies by MAIPA reveals an immunogenicity similar to the HPA-5 alloantigens.

Berry JE, Murphy CM, Smith GA, Ranasinghe E, Finberg R, Walton J, Brown J, Navarrete C, Metcalfe P, Ouwehand WH.

Division of Haematology, National Institute for Biological Standards and Control, Potters Bar, UK.

The glycosylphosphatidylinositol-linked platelet protein CD109 carries the biallelic alloantigen system Gov. There is limited information on the incidence of Gov alloantibodies in neonatal alloimmune thrombocytopenia (NAITP), post-transfusion purpura (PTP) and platelet refractoriness. We adapted the monoclonal antibody-specific immobilization of platelet antigens (MAIPA) assay to the detection of Gov antibodies and determined their incidence in 605 archived samples (112 with HPA antibodies) referred for the aforementioned conditions. Here, we show that CD109 expression was reduced upon platelet storage in saline or by cryopreservation, but was stable when stored as whole blood or therapeutic platelet concentrate. Fourteen of the 605 samples contained Gov alloantibodies (anti-Gova, n = 10; anti-Govb, n = 4), with the majority in platelet refractoriness (n = 9) and, of the remaining five, four in NAITP and one in PTP. In seven cases, no other HPA antibodies were detected, three being NAITP cases. The incidence of Gov antibodies was significantly lower than HPA-1 system antibodies (n = 87), but equalled the number of HPA-5 system antibodies (n = 14) and outnumbered HPA-2 and -3 system antibodies (10 altogether).

PMID: 10997989 [PubMed - indexed for MEDLINE]

May 12 2004 06:43:50

The Sequence of the Human Genome

J. Craig Venter,[1]* Mark D. Adams,[1] Eugene W. Myers,[1] Peter W. Li,[1] Richard J. Mural,[1] Granger G. Sutton,[1] Hamilton O. Smith,[1] Mark Yandell,[1] Cheryl A. Evans,[1] Robert A. Holt,[1] Jeannine D. Gocayne,[1] Peter Amanatides,[1] Richard M. Ballew,[1] Daniel H. Huson,[1] Jennifer Russo Wortman,[1] Qing Zhang,[1] Chinnappa D. Kodira,[1] Xiangqun H. Zheng,[1] Lin Chen,[1] Marian Skupski,[1] Gangadharan Subramanian,[1] Paul D. Thomas,[1] Jinghui Zhang,[1] George L. Gabor Miklos,[2] Catherine Nelson,[3] Samuel Broder,[1] Andrew G. Clark,[4] Joe Nadeau,[5] Victor A. McKusick,[6] Norton Zinder,[7] Arnold J. Levine,[7] Richard J. Roberts,[8] Mel Simon,[9] Carolyn Slayman,[10] Michael Hunkapiller,[11] Randall Bolanos,[1] Arthur Delcher,[1] Ian Dew,[1] Daniel Fasulo,[1] Michael Flanigan,[1] Liliana Florea,[1] Aaron Halpern,[1] Sridhar Hannenhalli,[1] Saul Kravitz,[1] Samuel Levy,[1] Clark Mobarry,[1] Knut Reinert,[1] Karin Remington,[1] Jane Abu-Threideh,[1] Ellen Beasley,[1] Kendra Biddick,[1] Vivien Bonazzi,[1] Rhonda Brandon,[1] Michele Cargill,[1] Ishwar Chandramouliswaran,[1] Rosane Charlab,[1] Kabir Chaturvedi,[1] Zuoming Deng,[1] Valentina Di Francesco,[1] Patrick Dunn,[1] Karen Eilbeck,[1] Carlos Evangelista,[1] Andrei E. Gabrielian,[1] Weiniu Gan,[1] Wangmao Ge,[1] Fangcheng Gong,[1] Zhiping Gu,[1] Ping Guan,[1] Thomas J. Heiman,[1] Maureen E. Higgins,[1] Rui-Ru Ji,[1] Zhaoxi Ke,[1] Karen A. Ketchum,[1] Zhongwu Lai,[1] Yiding Lei,[1] Zhenya Li,[1] Jiayin Li,[1] Yong Liang,[1] Xiaoying Lin,[1] Fu Lu,[1] Gennady V. Merkulov,[1] Natalia Milshina,[1] Helen M. Moore,[1] Ashwinikumar K Naik,[1] Vaibhav A. Narayan,[1] Beena Neelam,[1] Deborah Nusskern,[1] Douglas B. Rusch,[1] Steven Salzberg,[12] Wei Shao,[1] Bixiong Shue,[1] Jingtao Sun,[1] Zhen Yuan Wang,[1] Aihui Wang,[1] Xin Wang,[1] Jian Wang,[1] Ming-Hui Wei,[1] Ron Wides,[13] Chunlin Xiao,[1] Chunhua Yan,[1] Alison Yao,[1] Jane Ye,[1] Ming Zhan,[1] Weiqing Zhang,[1] Hongyu Zhang,[1] Qi Zhao,[1] Liansheng Zheng,[1] Fei Zhong,[1] Wenyan Zhong,[1] Shiaoping C. Zhu,[1] Shaying Zhao,[12] Dennis Gilbert,[1] Suzanna Baumhueter,[1] Gene Spier,[1] Christine Carter,[1] Anibal Cravchik,[1] Trevor Woodage,[1] Feroze Ali,[1] Huijin An,[1] Aderonke Awe,[1] Danita Baldwin,[1] Holly Baden,[1] Mary Barnstead,[1] Ian Barrow,[1] Karen Beeson,[1] Dana Busam,[1] Amy Carver,[1] Angela Center,[1] Ming Lai Cheng,[1] Liz Curry,[1] Steve Danaher,[1] Lionel Davenport,[1] Raymond Desilets,[1] Susanne Dietz,[1] Kristina Dodson,[1] Lisa Doup,[1] Steven Ferriera,[1] Neha Garg,[1] Andres Gluecksmann,[1] Brit Hart,[1] Jason Haynes,[1] Charles Haynes,[1] Cheryl Heiner,[1] Suzanne Hladun,[1] Damon Hostin,[1] Jarrett Houck,[1] Timothy Howland,[1] Chinyere Ibegwam,[1] Jeffery Johnson,[1] Francis Kalush,[1] Lesley Kline,[1] Shashi Koduru,[1] Amy Love,[1] Felecia Mann,[1] David May,[1] Steven McCawley,[1] Tina McIntosh,[1] Ivy McMullen,[1] Mee Moy,[1] Linda Moy,[1] Brian Murphy,[1] Keith Nelson,[1] Cynthia Pfannkoch,[1] Eric Pratts,[1] Vinita Puri,[1] Hina Qureshi,[1] Matthew Reardon,[1] Robert Rodriguez,[1] Yu-Hui Rogers,[1] Deanna Romblad,[1] Bob Ruhfel,[1] Richard Scott,[1] Cynthia Sitter,[1] Michelle Smallwood,[1] Erin Stewart,[1] Renee Strong,[1] Ellen Suh,[1] Reginald Thomas,[1] Ni Ni Tint,[1] Sukyee Tse,[1] Claire Vech,[1] Gary Wang,[1] Jeremy Wetter,[1] Sherita Williams,[1] Monica Williams,[1] Sandra Windsor,[1] Emily Winn-Deen,[1] Keriellen Wolfe,[1] Jayshree Zaveri,[1] Karena Zaveri,[1] Josep F. Abril,[14] Roderic Guigó,[14] Michael J. Campbell,[1] Kimmen V. Sjolander,[1] Brian Karlak,[1] Anish Kejariwal,[1] Huaiyu Mi,[1] Betty Lazareva,[1] Thomas Hatton,[1] Apurva Narechania,[1] Karen Diemer,[1] Anushya Muruganujan,[1] Nan Guo,[1] Shinji Sato,[1] Vineet Bafna,[1] Sorin Istrail,[1] Ross Lippert,[1] Russell Schwartz,[1] Brian Walenz,[1] Shibu Yooseph,[1] David Allen,[1] Anand Basu,[1] James Baxendale,[1] Louis Blick,[1] Marcelo Caminha,[1] John Carnes-Stine,[1] Parris Caulk,[1] Yen-Hui Chiang,[1] My Coyne,[1] Carl Dahlke,[1] Anne Deslattes Mays,[1] Maria Dombroski,[1] Michael Donnelly,[1] Dale Ely,[1] Shiva Esparham,[1] Carl Fosler,[1] Harold Gire,[1] Stephen Glanowski,[1] Kenneth Glasser,[1] Anna Glodek,[1] Mark Gorokhov,[1] Ken Graham,[1] Barry Gropman,[1] Michael Harris,[1] Jeremy Heil,[1] Scott Henderson,[1] Jeffrey Hoover,[1] Donald Jennings,[1] Catherine Jordan,[1] James Jordan,[1] John Kasha,[1] Leonid Kagan,[1] Cheryl Kraft,[1] Alexander Levitsky,[1] Mark Lewis,[1] Xiangjun Liu,[1] John Lopez,[1] Daniel Ma,[1] William Majoros,[1] Joe McDaniel,[1] Sean Murphy,[1] Matthew Newman,[1] Trung Nguyen,[1] Ngoc Nguyen,[1] Marc Nodell,[1] Sue Pan,[1] Jim Peck,[1] Marshall Peterson,[1] William Rowe,[1] Robert Sanders,[1] John Scott,[1] Michael Simpson,[1] Thomas Smith,[1] Arlan Sprague,[1] Timothy Stockwell,[1] Russell Turner,[1] Eli Venter,[1] Mei Wang,[1] Meiyuan Wen,[1] David Wu,[1] Mitchell Wu,[1] Ashley Xia,[1] Ali Zandieh,[1] Xiaohong Zhu[1]

A 2.91-billion base pair (bp) consensus sequence of the euchromatic portion of the human genome was generated by the whole-genome shotgun sequencing method. The 14.8-billion bp DNA sequence was generated over 9 months from 27,271,853 high-quality sequence reads (5.11-fold coverage of the genome) from both ends of plasmid clones made from the DNA of five individuals. Two assembly strategies—a whole-genome assembly and a regional chromosome assembly—were used, each combining sequence data from Celera and the publicly funded genome effort. The public data were shredded into 550-bp segments to create a 2.9-fold coverage of those genome regions that had been sequenced, without including biases inherent in the cloning and assembly procedure used by the publicly funded group. This brought the effective coverage in the assemblies to eightfold, reducing the number and size of gaps in the final assembly over what would be obtained with 5.11-fold coverage. The two assembly strategies yielded very similar results that largely agree with independent mapping data. The assemblies effectively cover the euchromatic regions of the human chromosomes. More than 90% of the genome is in scaffold assemblies of 100,000 bp or more, and 25% of the genome is in scaffolds of 10 million bp or larger. Analysis of the genome sequence revealed 26,588 protein-encoding transcripts for which there was strong corroborating evidence and an additional ~12,000 computationally derived genes with mouse matches or other weak supporting evidence. Although gene-dense clusters are obvious, almost half the genes are dispersed in low G+C sequence separated by large tracts of apparently noncoding sequence. Only 1.1% of the genome is spanned by exons, whereas 24% is in introns, with 75% of the genome being intergenic DNA. Duplications of segmental blocks, ranging in size up to chromosomal lengths, are abundant throughout the genome and reveal a complex evolutionary history. Comparative genomic analysis indicates vertebrate expansions of genes associated with neuronal function, with tissue-specific developmental regulation, and with the hemostasis and immune systems. DNA sequence comparisons between the consensus sequence and publicly funded genome data provided locations of 2.1 million single-nucleotide polymorphisms (SNPs). A random pair of human haploid genomes differed at a rate of 1 bp per 1250 on average, but there was marked heterogeneity in the level of polymorphism across the genome. Less than 1% of all SNPs resulted in variation in proteins, but the task of determining which SNPs have functional consequences remains an open challenge.

Decoding of the DNA that constitutes the human genome has been widely anticipated for the contribution it will make toward understanding human evolution, the causation of disease, and the interplay between the environment and heredity in defining the human condition. A project with the goal of determining the complete nucleotide sequence of the human genome was first formally proposed in 1985 (*1*). In subsequent years, the idea met with mixed reactions in the scientific community (*2*). However, in 1990, the Human Genome Project (HGP) was officially initiated in the United States under the direction of the National Institutes of Health and the U.S. Department of Energy with a 15-year, $3 billion plan for completing the genome sequence. In 1998 we announced our intention to build a unique genome-sequencing facility, to determine the sequence of the human genome over a 3-year period. Here we report the penultimate milestone along the path toward that goal, a nearly complete sequence of the euchromatic portion of the human genome. The sequencing was performed by a whole-genome random shotgun method with subsequent assembly of the sequenced segments.

The modern history of DNA sequencing began in 1977, when Sanger reported his method for determining the order of nucleotides of DNA using chain-terminating nucleotide analogs (*3*). In the same year, the first human gene was isolated and sequenced (*4*). In 1986, Hood and co-workers (*5*) described an improvement in the Sanger sequencing method that included attaching fluorescent dyes to the nucleotides, which permitted them to be sequentially read by a computer. The first automated DNA sequencer, developed by Applied Biosystems in California in 1987, was shown to be successful when the sequences of two genes were obtained with this new technology (*6*). From early sequencing of human genomic regions (*7*), it became clear that cDNA sequences (which are reverse-transcribed from RNA) would be essential to annotate and validate gene predictions in the human genome. These studies were the basis in part for the development of the expressed sequence tag (EST) method of gene identification (*8*), which is a random selection, very high throughput sequencing approach to characterize cDNA libraries. The EST method led to the rapid discovery and mapping of human genes (*9*). The increasing numbers of human EST sequences necessitated the development of new computer algorithms to analyze large amounts of sequence data, and in 1993 at The Institute for Genomic Research (TIGR), an algorithm was developed that permitted assembly and analysis of hundreds of thousands of ESTs. This algorithm permitted characterization and annotation of human genes on the basis of 30,000 EST assemblies (*10*).

The complete 49-kbp bacteriophage lambda genome sequence was determined by a shotgun restriction digest method in 1982 (*11*). When considering methods for sequencing the smallpox virus genome in 1991 (*12*), a whole-genome shotgun sequencing method was discussed and subsequently rejected owing to the lack of appropriate software tools for genome assembly. However, in 1994, when a microbial genome-sequencing project was contemplated at TIGR, a whole-genome shotgun sequencing approach was considered possible with the TIGR EST assembly algorithm. In 1995, the 1.8-Mbp *Haemophilus influenzae* genome was completed by a whole-genome shotgun sequencing method (*13*). The experience with several subsequent genome-sequencing efforts established the broad applicability of this approach (*14*, *15*).

A key feature of the sequencing approach used for these megabase-size and larger genomes was the use of paired-end sequences (also called mate pairs), derived from subclone libraries with distinct insert sizes and cloning characteristics. Paired-end sequences are sequences 500 to 600 bp in length from both ends of double-stranded DNA clones of prescribed lengths. The success of using end sequences from long segments (18 to 20 kbp) of DNA cloned into bacteriophage lambda in assembly of the microbial genomes led to the suggestion (*16*) of an approach to simulta-

[1]Celera Genomics, 45 West Gude Drive, Rockville, MD 20850, USA. [2]GenetixXpress, 78 Pacific Road, Palm Beach, Sydney 2108, Australia. [3]Berkeley *Drosophila* Genome Project, University of California, Berkeley, CA 94720, USA. [4]Department of Biology, Penn State University, 208 Mueller Lab, University Park, PA 16802, USA. [5]Department of Genetics, Case Western Reserve University School of Medicine, BRB-630, 10900 Euclid Avenue, Cleveland, OH 44106, USA. [6]Johns Hopkins University School of Medicine, Johns Hopkins Hospital, 600 North Wolfe Street, Blalock 1007, Baltimore, MD 21287–4922, USA. [7]Rockefeller University, 1230 York Avenue, New York, NY 10021–6399, USA. [8]New England BioLabs, 32 Tozer Road, Beverly, MA 01915, USA. [9]Division of Biology, 147-75, California Institute of Technology, 1200 East California Boulevard, Pasadena, CA 91125, USA. [10]Yale University School of Medicine, 333 Cedar Street, P.O. Box 208000, New Haven, CT 06520–8000, USA. [11]Applied Biosystems, 850 Lincoln Centre Drive, Foster City, CA 94404, USA. [12]The Institute for Genomic Research, 9712 Medical Center Drive, Rockville, MD 20850, USA. [13]Faculty of Life Sciences, Bar-Ilan University, Ramat-Gan, 52900 Israel. [14]Grup de Recerca en Informàtica Mèdica, Institut Municipal d'Investigació Mèdica, Universitat Pompeu Fabra, 08003-Barcelona, Catalonia, Spain.

*To whom correspondence should be addressed. E-mail: humangenome@celera.com

neously map and sequence the human genome by means of end sequences from 150-kbp bacterial artificial chromosomes (BACs) (*17*, *18*). The end sequences spanned by known distances provide long-range continuity across the genome. A modification of the BAC end-sequencing (BES) method was applied successfully to complete chromosome 2 from the *Arabidopsis thaliana* genome (*19*).

In 1997, Weber and Myers (*20*) proposed whole-genome shotgun sequencing of the human genome. Their proposal was not well received (*21*). However, by early 1998, as less than 5% of the genome had been sequenced, it was clear that the rate of progress in human genome sequencing worldwide was very slow (*22*), and the prospects for finishing the genome by the 2005 goal were uncertain.

In early 1998, PE Biosystems (now Applied Biosystems) developed an automated, high-throughput capillary DNA sequencer, subsequently called the ABI PRISM 3700 DNA Analyzer. Discussions between PE Biosystems and TIGR scientists resulted in a plan to undertake the sequencing of the human genome with the 3700 DNA Analyzer and the whole-genome shotgun sequencing techniques developed at TIGR (*23*). Many of the principles of operation of a genome-sequencing facility were established in the TIGR facility (*24*). However, the facility envisioned for Celera would have a capacity roughly 50 times that of TIGR, and thus new developments were required for sample preparation and tracking and for whole-genome assembly. Some argued that the required 150-fold scale-up from the *H. influenzae* genome to the human genome with its complex repeat sequences was not feasible (*25*). The *Drosophila melanogaster* genome was thus chosen as a test case for whole-genome assembly on a large and complex eukaryotic genome. In collaboration with Gerald Rubin and the Berkeley *Drosophila* Genome Project, the nucleotide sequence of the 120-Mbp euchromatic portion of the *Drosophila* genome was determined over a 1-year period (*26*–*28*). The *Drosophila* genome-sequencing effort resulted in two key findings: (i) that the assembly algorithms could generate chromosome assemblies with highly accurate order and orientation with substantially less than 10-fold coverage, and (ii) that undertaking multiple interim assemblies in place of one comprehensive final assembly was not of value.

These findings, together with the dramatic changes in the public genome effort subsequent to the formation of Celera (*29*), led to a modified whole-genome shotgun sequencing approach to the human genome. We initially proposed to do 10-fold sequence coverage of the genome over a 3-year period and to make interim assembled sequence data available quarterly. The modifications included a plan to perform random shotgun sequencing to ~5-fold coverage and to use the unordered and unoriented BAC sequence fragments and subassemblies published in GenBank by the publicly funded genome effort (*30*) to accelerate the project. We also abandoned the quarterly announcements in the absence of interim assemblies to report.

Although this strategy provided a reasonable result very early that was consistent with a whole-genome shotgun assembly with eight-fold coverage, the human genome sequence is not as finished as the *Drosophila* genome was with an effective 13-fold coverage. However, it became clear that even with this reduced coverage strategy, Celera could generate an accurately ordered and oriented scaffold sequence of the human genome in less than 1 year. Human genome sequencing was initiated 8 September 1999 and completed 17 June 2000. The first assembly was completed 25 June 2000, and the assembly reported here was completed 1 October 2000. Here we describe the whole-genome random shotgun sequencing effort applied to the human genome. We developed two different assembly approaches for assembling the ~3 billion bp that make up the 23 pairs of chromosomes of the *Homo sapiens* genome. Any GenBank-derived data were shredded to remove potential bias to the final sequence from chimeric clones, foreign DNA contamination, or misassembled contigs. Insofar as a correctly and accurately assembled genome sequence with faithful order and orientation of contigs is essential for an accurate analysis of the human genetic code, we have devoted a considerable portion of this manuscript to the documentation of the quality of our reconstruction of the genome. We also describe our preliminary analysis of the human genetic code on the basis of computational methods. Figure 1 (see fold-out chart associated with this issue; files for each chromosome can be found in Web fig. 1 on *Science* Online at www.sciencemag.org/cgi/content/full/291/5507/1304/DC1) provides a graphical overview of the genome and the features encoded in it. The detailed manual curation and interpretation of the genome are just beginning.

To aid the reader in locating specific analytical sections, we have divided the paper into seven broad sections. A summary of the major results appears at the beginning of each section.

1 Sources of DNA and Sequencing Methods
2 Genome Assembly Strategy and Characterization
3 Gene Prediction and Annotation
4 Genome Structure
5 Genome Evolution
6 A Genome-Wide Examination of Sequence Variations
7 An Overview of the Predicted Protein-Coding Genes in the Human Genome
8 Conclusions

1 Sources of DNA and Sequencing Methods

Summary. This section discusses the rationale and ethical rules governing donor selection to ensure ethnic and gender diversity along with the methodologies for DNA extraction and library construction. The plasmid library construction is the first critical step in shotgun sequencing. If the DNA libraries are not uniform in size, nonchimeric, and do not randomly represent the genome, then the subsequent steps cannot accurately reconstruct the genome sequence. We used automated high-throughput DNA sequencing and the computational infrastructure to enable efficient tracking of enormous amounts of sequence information (27.3 million sequence reads; 14.9 billion bp of sequence). Sequencing and tracking from both ends of plasmid clones from 2-, 10-, and 50-kbp libraries were essential to the computational reconstruction of the genome. Our evidence indicates that the accurate pairing rate of end sequences was greater than 98%.

Various policies of the United States and the World Medical Association, specifically the Declaration of Helsinki, offer recommendations for conducting experiments with human subjects. We convened an Institutional Review Board (IRB) (*31*) that helped us establish the protocol for obtaining and using human DNA and the informed consent process used to enroll research volunteers for the DNA-sequencing studies reported here. We adopted several steps and procedures to protect the privacy rights and confidentiality of the research subjects (donors). These included a two-stage consent process, a secure random alphanumeric coding system for specimens and records, circumscribed contact with the subjects by researchers, and options for off-site contact of donors. In addition, Celera applied for and received a Certificate of Confidentiality from the Department of Health and Human Services. This Certificate authorized Celera to protect the privacy of the individuals who volunteered to be donors as provided in Section 301(d) of the Public Health Service Act 42 U.S.C. 241(d).

Celera and the IRB believed that the initial version of a completed human genome should be a composite derived from multiple donors of diverse ethnic backgrounds. Prospective donors were asked, on a voluntary basis, to self-designate an ethnogeographic category (e.g., African-American, Chinese, Hispanic, Caucasian, etc.). We enrolled 21 donors (*32*).

Three basic items of information from each donor were recorded and linked by confidential code to the donated sample: age, sex, and self-designated ethnogeographic group. From females, ~130 ml of whole, heparinized blood was collected. From males, ~130 ml of whole, heparinized blood was

collected, as well as five specimens of se[illegible] collected over a 6-week period. Permanent lymphoblastoid cell lines were created by Epstein-Barr virus immortalization. DNA from five subjects was selected for genomic DNA sequencing: two males and three females—one African-American, one Asian-Chinese, one Hispanic-Mexican, and two Caucasians (see Web fig. 2 on *Science* Online at www.sciencemag.org/cgi/content/291/5507/1304/DC1). The decision of whose DNA to sequence was based on a complex mix of factors, including the goal of achieving diversity as well as technical issues such as the quality of the DNA libraries and availability of immortalized cell lines.

1.1 Library construction and sequencing

Central to the whole-genome shotgun sequencing process is preparation of high-quality plasmid libraries in a variety of insert sizes so that pairs of sequence reads (mates) are obtained, one read from both ends of each plasmid insert. High-quality libraries have an equal representation of all parts of the genome, a small number of clones without inserts, and no contamination from such sources as the mitochondrial genome and *Escherichia coli* genomic DNA. DNA from each donor was used to construct plasmid libraries in one or more of three size classes: 2 kbp, 10 kbp, and 50 kbp (Table 1) (*33*).

In designing the DNA-sequencing process, we focused on developing a simple system that could be implemented in a robust and reproducible manner and monitored effectively (Fig. 2) (*34*).

Current sequencing protocols are based on the dideoxy sequencing method (*35*), which typically yields only 500 to 750 bp of sequence per reaction. This limitation on read length has made monumental gains in throughput a prerequisite for the analysis of large eukaryotic genomes. We accomplished this at the Celera facility, which occupies about 30,000 square feet of laboratory space and produces sequence data continuously at a rate of 175,000 total reads per day. The DNA-sequencing facility is supported by a high-performance computational facility (*36*).

The process for DNA sequencing was modular by design and automated. Intermodule sample backlogs allowed four principal modules to operate independently: (i) library transformation, plating, and colony picking; (ii) DNA template preparation; (iii) dideoxy sequencing reaction set-up and purification; and (iv) sequence determination with the ABI PRISM 3700 DNA Analyzer. Because the inputs and outputs of each module have been carefully matched and sample backlogs are continuously managed, sequencing has proceeded without a single day's interruption since the initiation of the *Drosophila* project in May 1999. The ABI 3700 is a fully automated capillary array sequencer and as such can be operated with a minimal amount of hands-on time, currently estimated at about 15 min per day. The capillary system also facilitates correct associations of sequencing traces with samples through the elimination of manual sample loading and lane-tracking errors associated with slab gels. About 65 production staff were hired and trained, and were rotated on a regular basis through the four production modules. A central laboratory information management system (LIMS) tracked all sample plates by unique bar code identifiers. The facility was supported by a quality control team that performed raw material and in-process testing and a quality assurance group with responsibilities including document control, validation, and auditing of the facility. Critical to the success of the scale-up was the validation of all software and instrumentation before implementation, and production-scale testing of any process changes.

1.2 Trace processing

An automated trace-processing pipeline has been developed to process each sequence file (*37*). After quality and vector trimming, the average trimmed sequence length was 543 bp, and the sequencing accuracy was exponentially distributed with a mean of 99.5% and with less than 1 in 1000 reads being less than 98% accurate (*26*). Each trimmed sequence was screened for matches to contaminants including sequences of vector alone, *E. coli* genomic DNA, and human mitochondrial DNA. The entire read for any sequence with a significant match to a contaminant was discarded. A total of 713 reads matched *E. coli* genomic DNA and 2114 reads matched the human mitochondrial genome.

1.3 Quality assessment and control

The importance of the base-pair level accuracy of the sequence data increases as the size and repetitive nature of the genome to be sequenced increases. Each sequence read must be placed uniquely in the ge-

Table 1. Celera-generated data input into assembly.

	Individual	Number of reads for different insert libraries				Total number of base pairs
		2 kbp	10 kbp	50 kbp	Total	
No. of sequencing reads	A	0	0	2,767,357	2,767,357	1,502,674,851
	B	11,736,757	7,467,755	66,930	19,271,442	10,464,393,006
	C	853,819	881,290	0	1,735,109	942,164,187
	D	952,523	1,046,815	0	1,999,338	1,085,640,534
	F	0	1,498,607	0	1,498,607	813,743,601
	Total	13,543,099	10,894,467	2,834,287	27,271,853	14,808,616,179
Fold sequence coverage (2.9-Gb genome)	A	0	0	0.52	0.52	
	B	2.20	1.40	0.01	3.61	
	C	0.16	1.17	0	0.32	
	D	0.18	0.20	0	0.37	
	F	0	0.28	0	0.28	
	Total	2.54	2.04	0.53	5.11	
Fold clone coverage	A	0	0	18.39	18.39	
	B	2.96	11.26	0.44	14.67	
	C	0.22	1.33	0	1.54	
	D	0.24	1.58	0	1.82	
	F	0	2.26	0	2.26	
	Total	3.42	16.43	18.84	38.68	
Insert size* (mean)	Average	1,951 bp	10,800 bp	50,715 bp		
Insert size* (SD)	Average	6.10%	8.10%	14.90%		
% Mates†	Average	74.50	80.80	75.60		

*Insert size and SD are calculated from assembly of mates on contigs. †% Mates is based on laboratory tracking of sequencing runs.

nome, and even a modest error rate can reduce the effectiveness of assembly. In addition, maintaining the validity of mate-pair information is absolutely critical for the algorithms described below. Procedural controls were established for maintaining the validity of sequence mate-pairs as sequencing reactions proceeded through the process, including strict rules built into the LIMS. The accuracy of sequence data produced by the Celera process was validated in the course of the *Drosophila* genome project (*26*). By collecting data for the entire human genome in a single facility, we were able to ensure uniform quality standards and the cost advantages associated with automation, an economy of scale, and process consistency.

2 Genome Assembly Strategy and Characterization

Summary. We describe in this section the two approaches that we used to assemble the genome. One method involves the computational combination of all sequence reads with shredded data from GenBank to generate an independent, nonbiased view of the genome. The second approach involves clustering all of the fragments to a region or chromosome on the basis of mapping information. The clustered data were then shredded and subjected to computational assembly. Both approaches provided essentially the same reconstruction of assembled DNA sequence with proper order and orientation. The second method provided slightly greater sequence coverage (fewer gaps) and was the principal sequence used for the analysis phase. In addition, we document the completeness and correctness of this assembly process

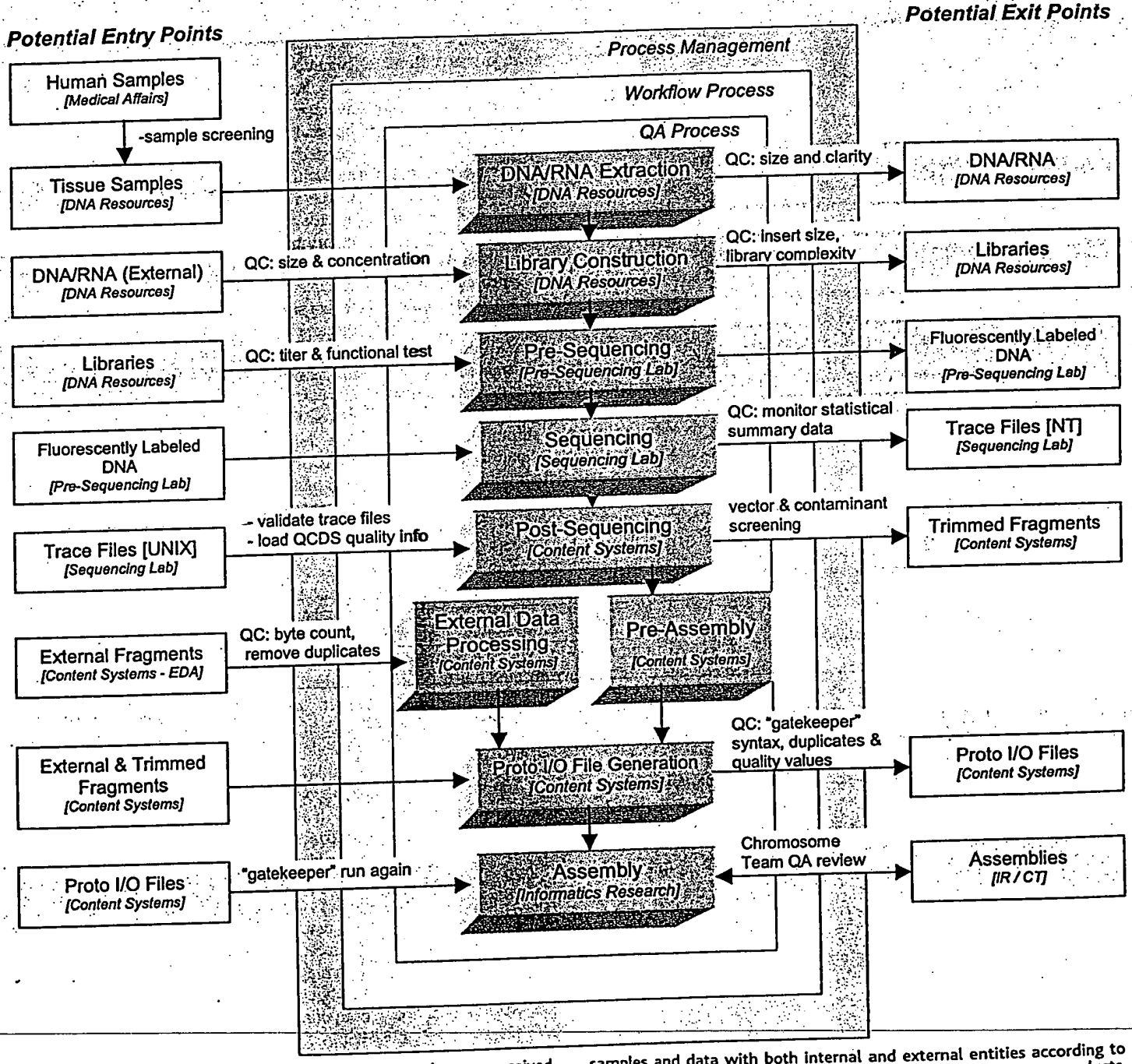


Fig. 2. Flow diagram for sequencing pipeline. Samples are received, selected, and processed in compliance with standard operating procedures, with a focus on quality within and across departments. Each process has defined inputs and outputs with the capability to exchange samples and data with both internal and external entities according to defined quality guidelines. Manufacturing pipeline processes, products, quality control measures, and responsible parties are indicated and are described further in the text.

and provide a comparison to the public gen[ome] sequence, which was reconstructed largely [by] an independent BAC-by-BAC approach. Our assemblies effectively covered the euchromatic regions of the human chromosomes. More than 90% of the genome was in scaffold assemblies of 100,000 bp or greater, and 25% of the genome was in scaffolds of 10 million bp or larger.

Shotgun sequence assembly is a classic example of an inverse problem: given a set of reads randomly sampled from a target sequence, reconstruct the order and the position of those reads in the target. Genome assembly algorithms developed for *Drosophila* have now been extended to assemble the ~25-fold larger human genome. Celera assemblies consist of a set of contigs that are ordered and oriented into scaffolds that are then mapped to chromosomal locations by using known markers. The contigs consist of a collection of overlapping sequence reads that provide a consensus reconstruction for a contiguous interval of the genome. Mate pairs are a central component of the assembly strategy. They are used to produce scaffolds in which the size of gaps between consecutive contigs is known with reasonable precision. This is accomplished by observing that a pair of reads, one of which is in one contig, and the other of which is in another, implies an orientation and distance between the two contigs (Fig. 3). Finally, our assemblies did not incorporate all reads into the final set of reported scaffolds. This set of unincorporated reads is termed "chaff," and typically consisted of reads from within highly repetitive regions, data from other organisms introduced through various routes as found in many genome projects, and data of poor quality or with untrimmed vector.

2.1 Assembly data sets

We used two independent sets of data for our assemblies. The first was a random shotgun data set of 27.27 million reads of average length 543 bp produced at Celera. This consisted largely of mate-pair reads from 16 libraries constructed from DNA samples taken from five different donors. Libraries with insert sizes of 2, 10, and 50 kbp were used. By looking at how mate pairs from a library were positioned in known sequenced stretches of the genome, we were able to characterize the range of insert sizes in each library and determine a mean and standard deviation. Table 1 details the number of reads, sequencing coverage, and clone coverage achieved by the data set. The clone coverage is the coverage of the genome in cloned DNA, considering the entire insert of each clone that has sequence from both ends. The clone coverage provides a measure of the amount of physical DNA coverage of the genome. Assuming a genome size of 2.9 Gbp, the Celera trimmed sequences gave a 5.1× coverage of the genome, and clone coverage was 3.42×, 16.40×, and 18.84× for the 2-, 10-, and 50-kbp libraries, respectively, for a total of 38.7× clone coverage.

The second data set was from the publicly funded Human Genome Project (PFP) and is primarily derived from BAC clones (*30*). The BAC data input to the assemblies came from a download of GenBank on 1 September 2000 (Table 2) totaling 4443.3 Mbp of sequence. The data for each BAC is deposited at one of four levels of completion. Phase 0 data are a set of generally unassembled sequencing reads from a very light shotgun of the BAC, typically less than 1×. Phase 1 data are unordered assemblies of contigs, which we call BAC contigs or bactigs. Phase 2 data are ordered assemblies of bactigs. Phase 3 data are complete BAC sequences. In the past 2 years the PFP has focused on a product of lower quality and completeness, but on a faster time-course, by concentrating on the production of Phase 1 data from a 3× to 4× light-shotgun of each BAC clone.

We screened the bactig sequences for contaminants by using the BLAST algorithm against three data sets: (i) vector sequences in Univec core (*38*), filtered for a 25-bp match at 98% sequence identity at the ends of the sequence and a 30-bp match internal to the sequence; (ii) the nonhuman portion of the High Throughput Genomic (HTG) Seqences division of GenBank (*39*), filtered at 200 bp at 98%; and (iii) the nonredundant nucleotide sequences from GenBank without primate and human virus entries, filtered at 200 bp at 98%. Whenever 25 bp or more of vector was found within 50 bp of the end of a contig, the tip up to the matching vector was excised. Under these criteria we removed 2.6 Mbp of possible contaminant and vector from the Phase 3 data, 61.0 Mbp from the Phase 1 and 2 data, and 16.1 Mbp from the Phase 0 data (Table 2). This left us with a total of 4363.7 Mbp of PFP sequence data 20% finished, 75% rough-draft (Phase 1 and 2), and 5% single sequencing reads (Phase 0). An additional 104,018 BAC end-sequence mate pairs were also downloaded and included in the data sets for both assembly processes (*18*).

2.2 Assembly strategies

Two different approaches to assembly were pursued. The first was a whole-genome assembly process that used Celera data and the PFP data in the form of additional synthetic shotgun data, and the second was a compartmentalized assembly process that first partitioned the Celera and PFP data into sets localized to large chromosomal segments and then performed ab initio shotgun assembly on each set. Figure 4 gives a schematic of the overall process flow.

For the whole-genome assembly, the PFP data was first disassembled or "shredded" into a synthetic shotgun data set of 550-bp reads that form a perfect 2× covering of the bactigs. This resulted in 16.05 million "faux" reads that were sufficient to cover the genome 2.96× because of redundancy in the BAC data set, without incorporating the biases inherent in the PFP assembly process. The combined data set of 43.32 million reads (8×), and all associated mate-pair information, were then subjected to our whole-genome assembly algorithm to produce a reconstruction of the genome. Neither the location of a BAC in the genome nor its assembly of bactigs was used in this process. Bactigs were shredded into reads because we found strong evidence that 2.13% of them were misassembled (*40*). Furthermore, BAC location

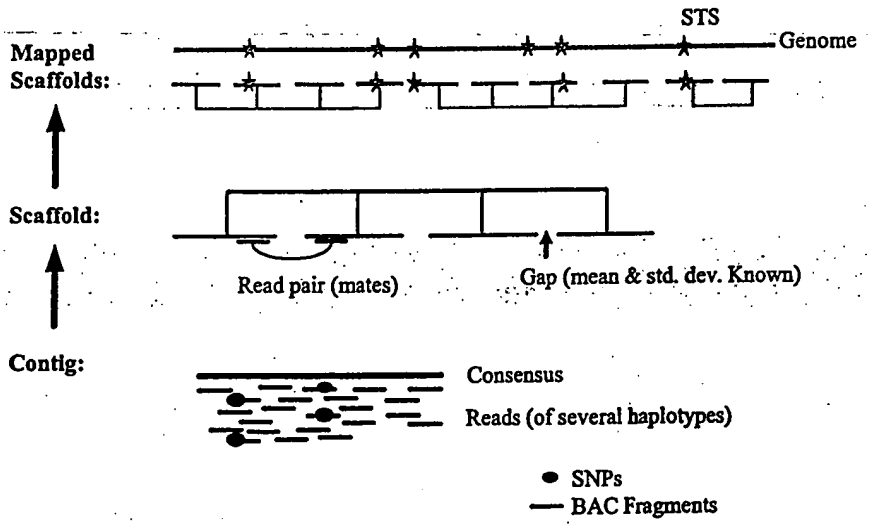


Fig. 3. Anatomy of whole-genome assembly. Overlapping shredded bactig fragments (red lines) and internally derived reads from five different individuals (black lines) are combined to produce a contig and a consensus sequence (green line). Contigs are connected into scaffolds (red) by using mate pair information. Scaffolds are then mapped to the genome (gray line) with STS (blue star) physical map information.

information was ignored because some BACs were not correctly placed on the PFP physical map and because we found strong evidence that at least 2.2% of the BACs contained sequence data that were not part of the given BAC (*41*), possibly as a result of sample-tracking errors (see below). In short, we performed a true, ab initio whole-genome assembly in which we took the expedient of deriving additional sequence coverage, but not mate pairs, assembled bactigs, or genome locality, from some externally generated data.

Table 2. GenBank data input into assembly.

Center	Statistics	Completion phase sequence		
		0	1 and 2	3
Whitehead Institute/ MIT Center for Genome Research, USA	Number of accession records	2,825	6,533	363
	Number of contigs	243,786	138,023	363
	Total base pairs	194,490,158	1,083,848,245	48,829,358
	Total vector masked (bp)	1,553,597	875,618	2,202
	Total contaminant masked (bp)	13,654,482	4,417,055	98,028
	Average contig length (bp)	798	7,853	134,516
Washington University, USA	Number of accession records	19	3,232	1,300
	Number of contigs	2,127	61,812	1,300
	Total base pairs	1,195,732	561,171,788	164,214,395
	Total vector masked (bp)	21,604	270,942	8,287
	Total contaminant masked (bp)	22,469	1,476,141	469,487
	Average contig length (bp)	562	9,079	126,319
Baylor College of Medicine, USA	Number of accession records	0	1,626	363
	Number of contigs	0	44,861	363
	Total base pairs	0	265,547,066	49,017,104
	Total vector masked (bp)	0	218,769	4,960
	Total contaminant masked (bp)	0	1,784,700	485,137
	Average contig length (bp)	0	5,919	135,033
Production Sequencing Facility, DOE Joint Genome Institute, USA	Number of accession records	135	2,043	754
	Number of contigs	7,052	34,938	754
	Total base pairs	8,680,214	294,249,631	60,975,328
	Total vector masked (bp)	22,644	162,651	7,274
	Total contaminant masked (bp)	665,818	4,642,372	118,387
	Average contig length (bp)	1,231	8,422	80,867
The Institute of Physical and Chemical Research (RIKEN), Japan	Number of accession records	0	1,149	300
	Number of contigs	0	25,772	300
	Total base pairs	0	182,812,275	20,093,926
	Total vector masked (bp)	0	203,792	2,371
	Total contaminant masked (bp)	0	308,426	27,781
	Average contig length (bp)	0	7,093	66,978
Sanger Centre, UK	Number of accession records	0	4,538	2,599
	Number of contigs	0	74,324	2,599
	Total base pairs	0	689,059,692	246,118,000
	Total vector masked (bp)	0	427,326	25,054
	Total contaminant masked (bp)	0	2,066,305	374,561
	Average contig length (bp)	0	9,271	94,697
Others*	Number of accession records	42	1,894	3,458
	Number of contigs	5,978	29,898	3,458
	Total base pairs	5,564,879	283,358,877	246,474,157
	Total vector masked (bp)	57,448	279,477	32,136
	Total contaminant masked (bp)	575,366	1,616,665	1,791,849
	Average contig length (bp)	931	9,478	71,277
All centers combined†	Number of accession records	3,021	21,015	9,137
	Number of contigs	258,943	409,628	9,137
	Total base pairs	209,930,983	3,360,047,574	835,722,268
	Total vector masked (bp)	1,655,293	2,438,575	82,284
	Total contaminant masked (bp)	14,918,135	16,311,664	3,365,230
	Average contig length (bp)	811	8,203	91,466

*Other centers contributing at least 0.1% of the sequence include: Chinese National Human Genome Center; Genomanalyse Gesellschaft fuer Biotechnologische Forschung mbH; Genome Therapeutics Corporation; GENOSCOPE; Chinese Academy of Sciences; Institute of Molecular Biotechnology; Keio University School of Medicine; Lawrence Livermore National Laboratory; Cold Spring Harbor Laboratory; Los Alamos National Laboratory; Max-Planck Institut fuer Molekulare, Genetik; Japan Science and Technology Corporation; Stanford University; The Institute for Genomic Research; The Institute of Physical and Chemical Research, Gene Bank; The University of Oklahoma; University of Texas Southwestern Medical Center, University of Washington. †The 4,405,700,825 bases contributed by all centers were shredded into faux reads resulting in 2.96× coverage of the genome.

In the compartmentalized shotgun assembly (CSA), Celera and PFP data were partitioned into the largest possible chromosomal segments or "components" that could be determined with confidence, and then shotgun assembly was applied to each partitioned subset wherein the bactig data were again shredded into faux reads to ensure an independent ab initio assembly of the component. By subsetting the data in this way, the overall computational effort was reduced and the effect of interchromosomal duplications was ameliorated. This also resulted in a reconstruction of the genome that was relatively independent of the whole-genome assembly results so that the two assemblies could be compared for consistency. The quality of the partitioning into components was crucial so that different genome regions were not mixed together. We constructed components from (i) the longest scaffolds of the sequence from each BAC and (ii) assembled scaffolds of data unique to Celera's data set. The BAC assemblies were obtained by a combining assembler that used the bactigs and the 5× Celera data mapped to those bactigs as input. This effort was undertaken as an interim step solely because the more accurate and complete the scaffold for a given sequence stretch, the more accurately one can tile these scaffolds into contiguous components on the basis of sequence overlap and mate-pair information. We further visually inspected and curated the scaffold tiling of the components to further increase its accuracy. For the final CSA assembly, all but the partitioning was ignored, and an independent, ab initio reconstruction of the sequence in each component was obtained by applying our whole-genome assembly algorithm to the partitioned, relevant Celera data and the shredded, faux reads of the partitioned, relevant bactig data.

2.3 Whole-genome assembly

The algorithms used for whole-genome assembly (WGA) of the human genome were enhancements to those used to produce the sequence of the *Drosophila* genome reported in detail in (*28*).

The WGA assembler consists of a pipeline composed of five principal stages: Screener, Overlapper, Unitigger, Scaffolder, and Repeat Resolver, respectively. The Screener finds and marks all microsatellite repeats with less than a 6-bp element, and screens out all known interspersed repeat elements, including Alu, Line, and ribosomal DNA. Marked regions get searched for overlaps, whereas screened regions do not get searched, but can be part of an overlap that involves unscreened matching segments.

The Overlapper compares every read against every other read in search of complete end-to-end overlaps of at least 40 bp and with no more than 6% differences in the match. Because all data are scrupulously vector-trimmed, the Overlapper can insist on complete overlap matches. Computing the set of all overlaps took roughly 10,000 CPU hours with a suite of four-processor Alpha SMPs with 4 gigabytes of RAM. This took 4 to 5 days in elapsed time with 40 such machines operating in parallel.

Every overlap computed above is statistically a 1-in-10^{17} event and thus not a coincidental event. What makes assembly combinatorially difficult is that while many overlaps are actually sampled from overlapping regions of the genome, and thus imply that the sequence reads should be assembled together, even more overlaps are actually from two distinct copies of a low-copy repeated element not screened above, thus constituting an error if put together. We call the former "true overlaps" and the latter "repeat-induced overlaps." The assembler must avoid choosing repeat-induced overlaps, especially early in the process.

We achieve this objective in the Unitigger. We first find all assemblies of reads that appear to be uncontested with respect to all other reads. We call the contigs formed from these subassemblies unitigs (for uniquely assembled contigs). Formally, these unitigs are the uncontested interval subgraphs of the graph of all overlaps (*42*). Unfortunately, although empirically many of these assemblies are correct (and thus involve only true overlaps), some are in fact collections of reads from several copies of a repetitive element that have been overcollapsed into a single subassembly. However, the overcollapsed unitigs are easily identified because their average coverage depth is too high to be consistent with the overall level of sequence coverage. We developed a simple statistical discriminator that gives the logarithm of the odds ratio that a unitig is composed of unique DNA or of a repeat consisting of two or more copies. The discriminator, set to a sufficiently stringent threshold, identifies a subset of the unitigs that we are certain are correct. In addition, a second, less stringent threshold identifies a subset of remaining unitigs very likely to be correctly assembled, of which we select those that will consistently scaffold (see below), and thus are again almost certain to be correct. We call the union of these two sets U-unitigs. Empirically, we found from a 6× simulated shotgun of human chromosome 22 that we get U-unitigs covering 98% of the stretches of unique DNA that are >2 kbp long. We are further able to identify the boundary of the start of a repetitive element at the ends of a U-unitig and leverage this so that U-unitigs span more than 93% of all singly interspersed Alu elements and other 100-to 400-bp repetitive segments.

The result of running the Unitigger was thus a set of correctly assembled subcontigs covering an estimated 73.6% of the human genome. The Scaffolder then proceeded to use mate-pair information to link these together into scaffolds. When there are two or more mate pairs that imply that a given pair of U-unitigs are at a certain distance and orientation with respect to each other, the probability of this being wrong is again roughly 1 in 10^{10}, assuming that mate pairs are false less than 2% of the time. Thus, one can with high confidence link together all U-unitigs that are linked by at least two 2- or 10-kbp mate pairs, producing intermediate-sized scaffolds that are then recursively linked together by confirming 50-kbp mate pairs and BAC end sequences. This process yielded scaffolds that are on the order of megabase pairs in size with gaps between their contigs that generally correspond to repetitive elements and occasionally to small sequencing gaps. These scaffolds reconstruct the majority of the unique sequence within a genome.

For the *Drosophila* assembly, we engaged in a three-stage repeat resolution strategy where each stage was progressively more aggressive and thus more likely to make a mistake. For the human assembly, we continued to use the first "Rocks" substage where all unitigs with a good, but not definitive, discriminator score are placed in a scaffold gap. This was done with the condition that two or more mate pairs with one of their reads already in the scaffold unambiguously place the unitig in the given gap. We estimate the probability of inserting a unitig into an incorrect gap with this strategy to be less than 10^{-7} based on a probabilistic analysis.

We revised the ensuing "Stones" substage of the human assembly, making it more like the mechanism suggested in our earlier work (*43*). For each gap, every read R that is placed in the gap by virtue of its mated pair M being in a contig of the scaffold and implying R's placement is collected. Celera's mate-pairing information is correct more than 99% of the time. Thus, almost every, but not all, of the reads in the set belong in the gap, and when a read does not belong it rarely agrees with the remainder of the reads. Therefore, we simply assemble this set of reads within the gap, eliminating any reads that conflict with the assembly. This operation proved much more reliable than the one it replaced for the *Drosophila* assembly; in the assembly of a simulated shotgun data set of human chromo-

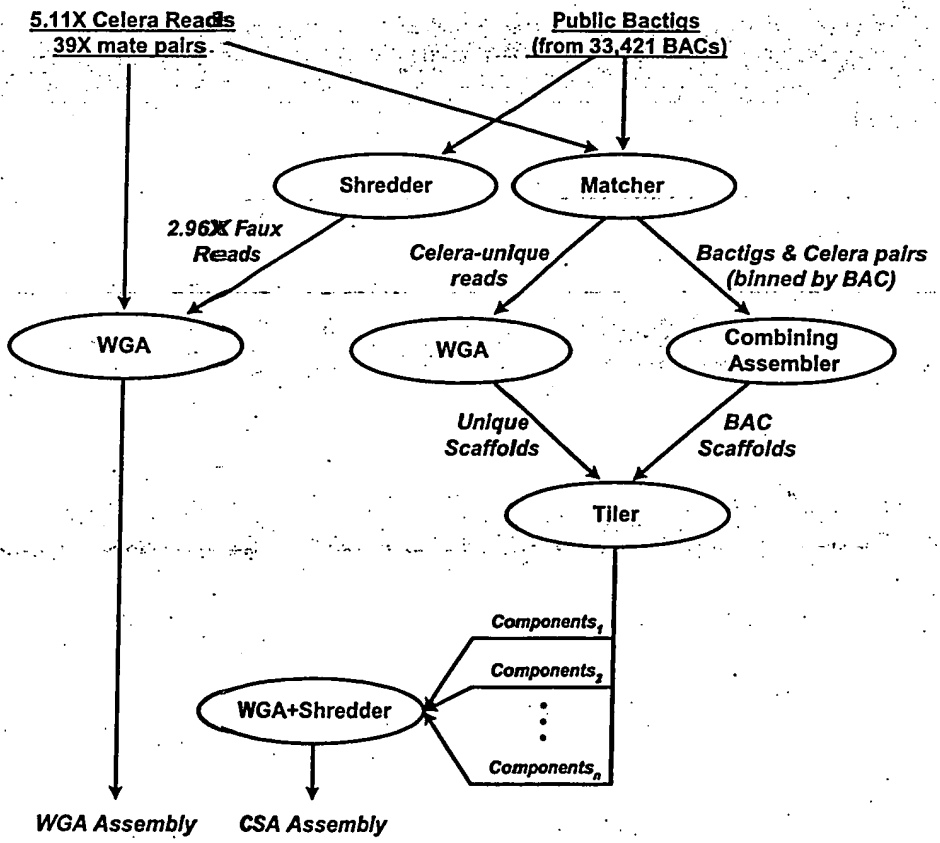


Fig. 4. Architecture of Celera's two-pronged assembly strategy. Each oval denotes a computation process performing the function indicated by its label, with the labels on arcs between ovals describing the nature of the objects produced and/or consumed by a process. This figure summarizes the discussion in the text that defines the terms and phrases used.

Sequence Revision History

Revision history for AC026605

This ID was replaced by AL590428 (See Rev. history)

GI	Version	Update Date	Status	I	II
8079076	3	May 26 2000 4:49 PM	Dead	◉	○
8079076	3	May 25 2000 1:14 PM	Dead	○	◉
7770540	2	May 12 2000 6:00 AM	Dead	○	○
7284630	1	Mar 22 2000 2:21 PM	Dead	○	○

Accession AC026605.3 was first seen at NCBI on Mar 22 2000 2:21 PM

Entrez

Search for Genes
LocusLink provides curated information for human, fruit fly, mouse, rat, and zebrafish

Help|FAQ

Batch Entrez: Upload a file of GI or accession numbers to retrieve protein or nucleotide sequences

Check sequence revision history

How to create WWW links to Entrez

LinkOut

Cubby

Related resources

BLAST

Reference sequence project

LocusLink

Clusters of orthologous groups

Protein reviews on the web

Query= SEQ ID NO:3
(4287 letters)

Sequences producing significant alignments:	Score (bits)	E Value
AL590428.7.1.163577	460	e-126
AL591480.8.1.91419	349	9e-93

>AL590428.7.1.163577
Length = 163577

Score = 460 bits (232), Expect = e-126
Identities = 232/232 (100%)
Strand = Plus / Plus

```
Query: 277    ctacctctgaacagtgcagatgagatttatgagctacgtgtaaccggacgtacccaggat 336
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 80550  ctacctctgaacagtgcagatgagatttatgagctacgtgtaaccggacgtacccaggat 80609


Query: 337    gagattttattctctaatagtacccgcttatcatttgagaccaagagaatatctgtcttc 396
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 80610  gagattttattctctaatagtacccgcttatcatttgagaccaagagaatatctgtcttc 80669


Query: 397    attcaaacagacaaggccttatacaagccaaagcaagaagtgaagtttcgcattgttaca 456
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 80670  attcaaacagacaaggccttatacaagccaaagcaagaagtgaagtttcgcattgttaca 80729


Query: 457    ctcttctcagattttaagccttacaaaacctctttaaacattctcattaagg 508
              ||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 80730  ctcttctcagattttaagccttacaaaacctctttaaacattctcattaagg 80781
```

Score = 456 bits (230), Expect = e-125
Identities = 230/230 (100%)
Strand = Plus / Plus

```
Query: 2960   ggttgtcagcttttgttttaagatgtttccttgaagccgatccttacatagatattgatc 3019
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 157049 ggttgtcagcttttgttttaagatgtttccttgaagccgatccttacatagatattgatc 157108


Query: 3020   agaatgtgttacacagaacatacacttggcttaaaggacatcagaaatccaacggtgaat 3079
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 157109 agaatgtgttacacagaacatacacttggcttaaaggacatcagaaatccaacggtgaat 157168


Query: 3080   tttgggatccaggaagagtgattcatagtgagcttcaaggtggcaataaaagtccagtaa 3139
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 157169 tttgggatccaggaagagtgattcatagtgagcttcaaggtggcaataaaagtccagtaa 157228
```

```
Query: 3140   cacttacagcctatattgtaacttctctcctgggatatagaaagtatcag 3189
              |||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 157229 cacttacagcctatattgtaacttctctcctgggatatagaaagtatcag 157278


 Score =  448 bits (226), Expect = e-122
 Identities = 226/226 (100%)
 Strand = Plus / Plus


Query: 1108   gtgaaggtaactcgtgctgatggcaaccaactgactcttgaagaaagaagaaataatgta 1167
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 116136 gtgaaggtaactcgtgctgatggcaaccaactgactcttgaagaaagaagaaataatgta 116195


Query: 1168   gtcataacagtgacacagagaaactatactgagtactggagcggatctaacagtggaaat 1227
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 116196 gtcataacagtgacacagagaaactatactgagtactggagcggatctaacagtggaaat 116255


Query: 1228   cagaaaatggaagctgttcagaaaataaattatactgtcccccaaagtggaacttttaag 1287
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 116256 cagaaaatggaagctgttcagaaaataaattatactgtcccccaaagtggaacttttaag 116315


Query: 1288   attgaattcccaatcctggaggattccagtgagctacagttgaagg 1333
              |||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 116316 attgaattcccaatcctggaggattccagtgagctacagttgaagg 116361


 Score =  432 bits (218), Expect = e-118
 Identities = 221/222 (99%)
 Strand = Plus / Plus


Query: 2336   aggttaaggtaatcattgagaaaagtgacaaatttgatattctaatgacttcaagtgaaa 2395
              |||||||||||||||||||||||||||||||||||||||||||||||||||||| |||||
Sbjct: 137438 aggttaaggtaatcattgagaaaagtgacaaatttgatattctaatgacttcaaatgaaa 137497


Query: 2396   taaatgccacaggccaccagcagacccttctggttcccagtgaggatggggcaactgttc 2455
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 137498 taaatgccacaggccaccagcagacccttctggttcccagtgaggatggggcaactgttc 137557


Query: 2456   tttttcccatcaggccaacacatctgggagaaattcctatcacagtcacagctctttcac 2515
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 137558 tttttcccatcaggccaacacatctgggagaaattcctatcacagtcacagctctttcac 137617


Query: 2516   ccactgcttctgatgctgtcacccagatgattttagtaaagg 2557
              |||||||||||||||||||||||||||||||||||||||||
Sbjct: 137618 ccactgcttctgatgctgtcacccagatgattttagtaaagg 137659
```

```
 Score =  385 bits (194), Expect = e-103
 Identities = 194/194 (100%)
 Strand = Plus / Plus


Query: 3354   aggtggcatgcaattctgggtgtcatcagagtccaaactttctgactcctggcagccacg 3413
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 160188 aggtggcatgcaattctgggtgtcatcagagtccaaactttctgactcctggcagccacg 160247


Query: 3414   ctccctggatattgaagttgcagcctatgcactgctctcacacttcttacaatttcagac 3473
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 160248 ctccctggatattgaagttgcagcctatgcactgctctcacacttcttacaatttcagac 160307


Query: 3474   ttctgagggaatcccaattatgaggtggctaagcaggcaaagaaatagcttgggtggttt 3533
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 160308 ttctgagggaatcccaattatgaggtggctaagcaggcaaagaaatagcttgggtggttt 160367


Query: 3534   tgcatctactcagg 3547
              ||||||||||||||
Sbjct: 160368 tgcatctactcagg 160381



 Score =  357 bits (180), Expect = 4e-95
 Identities = 180/180 (100%)
 Strand = Plus / Plus


Query: 2701   ggagatgttcttggtccttccatcaatggcttagcctcattgattcggatgccttatggc 2760
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 142831 ggagatgttcttggtccttccatcaatggcttagcctcattgattcggatgccttatggc 142890


Query: 2761   tgtggtgaacagaacatgataaattttgctccaaatatttacattttggattatctgact 2820
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 142891 tgtggtgaacagaacatgataaattttgctccaaatatttacattttggattatctgact 142950


Query: 2821   aaaaagaaacaactgacagataatttgaaagaaaaagctctttcatttatgaggcaaggt 2880
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 142951 aaaaagaaacaactgacagataatttgaaagaaaaagctctttcatttatgaggcaaggt 143010



 Score =  353 bits (178), Expect = 6e-94
 Identities = 178/178 (100%)
 Strand = Plus / Plus


Query: 1497   ggtagtatccaggggacagttggtggctgtaggaaaacaaaattcaacaatgttctcttt 1556
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 118260 ggtagtatccaggggacagttggtggctgtaggaaaacaaaattcaacaatgttctcttt 118319
```

```
Query: 1557   aacaccagaaaattcttggactccaaaagcctgtgtaattgtgtattatattgaagatga 1616
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 118320 aacaccagaaaattcttggactccaaaagcctgtgtaattgtgtattatattgaagatga 118379


Query: 1617   tggggaaattataagtgatgttctaaaaattcctgttcagcttgtttttaaaaataag 1674
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 118380 tggggaaattataagtgatgttctaaaaattcctgttcagcttgtttttaaaaataag 118437


 Score =  347 bits (175), Expect = 4e-92
 Identities = 175/175 (100%)
 Strand = Plus / Plus


Query: 74    ggcctcggtttctggtgacagccccagggatcatcaggcccggaggaaatgtgactattg 133
              |||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 47605 ggcctcggtttctggtgacagccccagggatcatcaggcccggaggaaatgtgactattg 47664


Query: 134   gggtggagcttctggaacactgcccttcacaggtgactgtgaaggcggagctgctcaaga 193
              |||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 47665 gggtggagcttctggaacactgcccttcacaggtgactgtgaaggcggagctgctcaaga 47724


Query: 194   cagcatcaaacctcactgtctctgtcctggaagcagaaggagtctttgaaaaagg 248
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 47725 cagcatcaaacctcactgtctctgtcctggaagcagaaggagtctttgaaaaagg 47779


 Score =  337 bits (170), Expect = 3e-89
 Identities = 170/170 (100%)
 Strand = Plus / Plus


Query: 3188   agcctaacattgatgtgcaagagtctatccatttttggagtctgaattcagtagaggaa 3247
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 158287 agcctaacattgatgtgcaagagtctatccatttttggagtctgaattcagtagaggaa 158346


Query: 3248   tttcagacaattatactctagcccttataacttatgcattgtcatcagtggggagtccta 3307
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 158347 tttcagacaattatactctagcccttataacttatgcattgtcatcagtggggagtccta 158406


Query: 3308   aagcgaaggaagctttgaatatgctgacttggagagcagaacaagaaggt 3357
              |||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 158407 aagcgaaggaagctttgaatatgctgacttggagagcagaacaagaaggt 158456
```

```
 Score =  313 bits (158), Expect = 5e-82
 Identities = 158/158 (100%)
 Strand = Plus / Plus


Query: 1673   agataaagctatattggagtaaagtgaaagctgaaccatctgagaaagtctctcttagga 1732
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 121633 agataaagctatattggagtaaagtgaaagctgaaccatctgagaaagtctctcttagga 121692


Query: 1733   tctctgtgacacagcctgactccatagttgggattgtagctgttgacaaaagtgtgaatc 1792
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 121693 tctctgtgacacagcctgactccatagttgggattgtagctgttgacaaaagtgtgaatc 121752


Query: 1793   tgatgaatgcctctaatgatattacaatggaaaatgtg 1830
              |||||||||||||||||||||||||||||||||||||
Sbjct: 121753 tgatgaatgcctctaatgatattacaatggaaaatgtg 121790



 Score =  293 bits (148), Expect = 5e-76
 Identities = 148/148 (100%)
 Strand = Plus / Plus


Query: 2555   aggctgaaggaatagaaaaatcatattcacaatccatcttattagacttgactgacaata 2614
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 138672 aggctgaaggaatagaaaaatcatattcacaatccatcttattagacttgactgacaata 138731


Query: 2615   ggctacagagtaccctgaaaactttgagtttctcatttcctcctaatacagtgactggca 2674
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 138732 ggctacagagtaccctgaaaactttgagtttctcatttcctcctaatacagtgactggca 138791


Query: 2675   gtgaaagagttcagatcactgcaattgg 2702
              ||||||||||||||||||||||||||||
Sbjct: 138792 gtgaaagagttcagatcactgcaattgg 138819



 Score =  289 bits (146), Expect = 7e-75
 Identities = 146/146 (100%)
 Strand = Plus / Plus


Query: 854    agataaatggatctgcaaacttctcttttaatgatgaagagatgaaaaatgtaatggatt 913
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 112945 agataaatggatctgcaaacttctcttttaatgatgaagagatgaaaaatgtaatggatt 113004


Query: 914    cttcaaatggactttctgaatacctggatctatcttcccctggaccagtagaaattttaa 973
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 113005 cttcaaatggactttctgaatacctggatctatcttcccctggaccagtagaaattttaa 113064
```

```
Query: 974    ccacagtgacagaatcagttacaggt 999
              |||||||||||||||||||||||||
Sbjct: 113065 ccacagtgacagaatcagttacaggt 113090
```

Score = 281 bits (142), Expect = 2e-72
Identities = 142/142 (100%)
Strand = Plus / Plus

```
Query: 1964   atgacaatgcagaatatgctgagaggtttatggaggaaaatgaaggacatattgtagata 2023
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 132820 atgacaatgcagaatatgctgagaggtttatggaggaaaatgaaggacatattgtagata 132879

Query: 2024   ttcatgacttttctttgggtagcagtccacatgtccgaaagcattttccagagacttgga 2083
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 132880 ttcatgacttttctttgggtagcagtccacatgtccgaaagcattttccagagacttgga 132939

Query: 2084   tttggctagacaccaacatggg 2105
              |||||||||||||||||||||
Sbjct: 132940 tttggctagacaccaacatggg 132961
```

Score = 256 bits (129), Expect = 1e-64
Identities = 129/129 (100%)
Strand = Plus / Plus

```
Query: 506   aggaccccaaatcaaatttgatccaacagtggttgtcacaacaaagtgatcttggagtca 565
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 86587 aggaccccaaatcaaatttgatccaacagtggttgtcacaacaaagtgatcttggagtca 86646

Query: 566   tttccaaaacttttcagctatcttcccatccaatacttggtgactggtctattcaagttc 625
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 86647 tttccaaaacttttcagctatcttcccatccaatacttggtgactggtctattcaagttc 86706

Query: 626   aagtgaatg 634
              |||||||||
Sbjct: 86707 aagtgaatg 86715
```

Score = 236 bits (119), Expect = 9e-59
Identities = 119/119 (100%)
Strand = Plus / Plus

```
Query: 2105   gttacaggatttaccaagaatttgaagtaactgtacctgattctatcacttcttgggtgg 2164
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 133912 gttacaggatttaccaagaatttgaagtaactgtacctgattctatcacttcttgggtgg 133971
```

```
Query: 2165   ctactggttttgtgatctctgaggacctgggtcttggactaacaactactccagtggag 2223
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 133972 ctactggttttgtgatctctgaggacctgggtcttggactaacaactactccagtggag 134030


 Score =  234 bits (118), Expect = 4e-58
 Identities = 118/118 (100%)
 Strand = Plus / Plus


Query: 2222   agctccaagccttccaaccatttttcatttttttgaatcttccctactctgttatcagag 2281
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 135568 agctccaagccttccaaccatttttcatttttttgaatcttccctactctgttatcagag 135627


Query: 2282   gtgaagaatttgctttggaaataactatattcaattatttgaaagatgccactgaggt 2339
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 135628 gtgaagaatttgctttggaaataactatattcaattatttgaaagatgccactgaggt 135685


 Score =  226 bits (114), Expect = 9e-56
 Identities = 114/114 (100%)
 Strand = Plus / Plus


Query: 996    aggtatttcaagaaatgtaagcactaatgtgttcttcaagcaacatgattacatcattga 1055
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 113780 aggtatttcaagaaatgtaagcactaatgtgttcttcaagcaacatgattacatcattga 113839


Query: 1056   gttttttgattatactactgtcttgaagccatctctcaacttcacagccactgt 1109
              ||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 113840 gttttttgattatactactgtcttgaagccatctctcaacttcacagccactgt 113893


 Score =  214 bits (108), Expect = 3e-52
 Identities = 108/108 (100%)
 Strand = Plus / Plus


Query: 3544   caggataccactgtggctttaaaggctctgtctgaatttgcagccctaatgaatacagaa 3603
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 161195 caggataccactgtggctttaaaggctctgtctgaatttgcagccctaatgaatacagaa 161254


Query: 3604   aggacaaatatccaagtgaccgtgacggggcctagctcaccaagtcct 3651
              |||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 161255 aggacaaatatccaagtgaccgtgacggggcctagctcaccaagtcct 161302
```

```
 Score =  210 bits (106), Expect = 5e-51
 Identities = 106/106 (100%)
 Strand = Plus / Plus


Query: 1331   aggcctatttccttggtagtaaaagtagcatggcagttcatagtctgtttaagtctccta 1390
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 116963 aggcctatttccttggtagtaaaagtagcatggcagttcatagtctgtttaagtctccta 117022


Query: 1391   gtaagacatacatccaactaaaaacaagagatgaaaatataaaggt 1436
              |||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 117023 gtaagacatacatccaactaaaaacaagagatgaaaatataaaggt 117068



 Score =  192 bits (97), Expect = 1e-45
 Identities = 97/97 (100%)
 Strand = Plus / Plus


Query: 759    gtatacatatgggaagccagtgaaaggagacgtaacgcttacatttttacctttatcctt 818
              |||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 112590 gtatacatatgggaagccagtgaaaggagacgtaacgcttacatttttacctttatcctt 112649


Query: 819    ttggggaaagaagaaaaatattacaaaaacatttaag 855
              |||||||||||||||||||||||||||||||||||||
Sbjct: 112650 ttggggaaagaagaaaaatattacaaaaacatttaag 112686



 Score =  176 bits (89), Expect = 8e-41
 Identities = 89/89 (100%)
 Strand = Plus / Plus


Query: 673    gtattaccaaaatttgaagtgactttgcagacaccattatattgttctatgaattctaag 732
              |||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 109149 gtattaccaaaatttgaagtgactttgcagacaccattatattgttctatgaattctaag 109208


Query: 733    catttaaatggtaccatcacggcaaagta 761
              |||||||||||||||||||||||||||||
Sbjct: 109209 catttaaatggtaccatcacggcaaagta 109237



 Score =  170 bits (86), Expect = 5e-39
 Identities = 86/86 (100%)
 Strand = Plus / Plus


Query: 2877   aggttaccagagagaacttctctatcagagggaagatggctctttcagtgcttttgggaa 2936
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 153424 aggttaccagagagaacttctctatcagagggaagatggctctttcagtgcttttgggaa 153483
```

```
Query: 2937   ttatgacccttctgggagcacttggt 2962
              |||||||||||||||||||||||||
Sbjct: 153484 ttatgacccttctgggagcacttggt 153509
```

Score = 153 bits (77), Expect = 1e-33
Identities = 80/81 (98%)
Strand = Plus / Plus

```
Query: 1828   gtggtccatgagttggaactttataacacaggatattatttaggcatgttcatgaattct 1887
              |||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 130630 gtggtccatgagttggaactttataacacaggatattatttaggcatgttcatgaattct 130689

Query: 1888   tttgcagtctttcaggaatgt 1908
              |||||||||||||||| ||||
Sbjct: 130690 tttgcagtctttcaggtatgt 130710
```

Score = 149 bits (75), Expect = 2e-32
Identities = 75/75 (100%)
Strand = Plus / Plus

```
Query: 1      atgcagggcccaccgctcctgaccgccgcccacctcctctgcgtgtgcaccgccgcgctg 60
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 46422  atgcagggcccaccgctcctgaccgccgcccacctcctctgcgtgtgcaccgccgcgctg 46481

Query: 61     gccgtggctcccggg 75
              |||||||||||||||
Sbjct: 46482  gccgtggctcccggg 46496
```

Score = 135 bits (68), Expect = 3e-28
Identities = 68/68 (100%)
Strand = Plus / Plus

```
Query: 1433   aggtgggatcgccttttgagttggtggttagtggcaacaaacgattgaaggagttaagct 1492
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 117152 aggtgggatcgccttttgagttggtggttagtggcaacaaacgattgaaggagttaagct 117211

Query: 1493   atatggta 1500
              ||||||||
Sbjct: 117212 atatggta 117219
```

```
 Score =  125 bits (63), Expect = 2e-25
 Identities = 63/63 (100%)
 Strand = Plus / Plus


Query: 1901   aggaatgtggactctgggtattgacagatgcaaacctcacgaaggattatattgatggtg 1960
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 131463 aggaatgtggactctgggtattgacagatgcaaacctcacgaaggattatattgatggtg 131522


Query: 1961   ttt 1963
              |||
Sbjct: 131523 ttt 131525



 Score =  123 bits (62), Expect = 1e-24
 Identities = 65/66 (98%)
 Strand = Plus / Plus


Query: 3652   cttgctgtggtacagccaatggcagttaatatttccgcaaatggttttggatttgctatt 3711
              |||||||||||||||||| ||||||||||||||||||||||||||||||||||||||||
Sbjct: 162411 cttgctgtggtacagccaacggcagttaatatttccgcaaatggttttggatttgctatt 162470


Query: 3712   tgtcag 3717
              ||||||
Sbjct: 162471 tgtcag 162476



 Score = 71.9 bits (36), Expect = 3e-09
 Identities = 39/40 (97%)
 Strand = Plus / Plus


Query: 634    gaccagacatattatcaatcatttcaggtttcagaatatg 673
              |||||||||||| ||||||||||||||||||||||||||
Sbjct: 106849 gaccagacatactatcaatcatttcaggtttcagaatatg 106888



 Score = 61.9 bits (31), Expect = 3e-06
 Identities = 31/31 (100%)
 Strand = Plus / Plus


Query: 246   aggctcttttaagacacttactcttccatca 276
              |||||||||||||||||||||||||||||
Sbjct: 73455 aggctcttttaagacacttactcttccatca 73485
```

>AL591480.8.1.91419
Length = 91419

Score = 349 bits (176), Expect = 9e-93
Identities = 176/176 (100%)
Strand = Plus / Plus

```
Query: 4112  ggagacaggcggtgagaagttacaactctgaagtgaagctgtcctcctgtgacctttgca 4171
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 12087 ggagacaggcggtgagaagttacaactctgaagtgaagctgtcctcctgtgacctttgca 12146


Query: 4172  gtgatgtccagggctgccgtccttgtgaggatggagcttcaggctcccatcatcactctt 4231
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 12147 gtgatgtccagggctgccgtccttgtgaggatggagcttcaggctcccatcatcactctt 12206


Query: 4232  cagtcatttttatttctgtttcaagcttctgtactttatggaactttggctgtga 4287
              ||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 12207 cagtcatttttatttctgtttcaagcttctgtactttatggaactttggctgtga 12262
```

Score = 305 bits (154), Expect = 1e-79
Identities = 154/154 (100%)
Strand = Plus / Plus

```
Query: 3859 agcttttcgggcccgggtaggagtggcatggctcttatggaagttaacctattaagtggc 3918
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 7015 agcttttcgggcccgggtaggagtggcatggctcttatggaagttaacctattaagtggc 7074


Query: 3919 tttatggtgccttcagaagcaatttctctgagcgagacagtgaagaaagtggaatatgat 3978
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 7075 tttatggtgccttcagaagcaatttctctgagcgagacagtgaagaaagtggaatatgat 7134


Query: 3979 catggaaaactcaacctctatttagattctgtaa 4012
              ||||||||||||||||||||||||||||||||
Sbjct: 7135 catggaaaactcaacctctatttagattctgtaa 7168
```

Score = 289 bits (146), Expect = 7e-75
Identities = 146/146 (100%)
Strand = Plus / Plus

```
Query: 3715 cagctcaatgttgtatataatgtgaaggcttctgggtcttctagaagacgaagatctatc 3774
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 3607 cagctcaatgttgtatataatgtgaaggcttctgggtcttctagaagacgaagatctatc 3666
```

```
Query: 3775 caaaatcaagaagcctttgatttagatgttgctgtaaaagaaaataaagatgatctcaat 3834
           ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 3667 caaaatcaagaagcctttgatttagatgttgctgtaaaagaaaataaagatgatctcaat 3726


Query: 3835 catgtggatttgaatgtgtgtacaag 3860
           |||||||||||||||||||||||||
Sbjct: 3727 catgtggatttgaatgtgtgtacaag 3752
```

Score = 206 bits (104), Expect = 8e-50
Identities = 104/104 (100%)
Strand = Plus / Plus

```
Query: 4009 gtaaatgaaacccagttttgtgttaatattcctgctgtgagaaactttaaagtttcaaat 4068
           ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 9090 gtaaatgaaacccagttttgtgttaatattcctgctgtgagaaactttaaagtttcaaat 9149


Query: 4069 acccaagatgcttcagtgtccatagtggattactatgagccaag 4112
           |||||||||||||||||||||||||||||||||||||||||||
Sbjct: 9150 acccaagatgcttcagtgtccatagtggattactatgagccaag 9193
```

Score = 131 bits (66), Expect = 4e-27
Identities = 66/66 (100%)
Strand = Plus / Plus

```
Query: 3652 cttgctgtggtacagccaatggcagttaatatttccgcaaatggttttggatttgctatt 3711
           ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 834  cttgctgtggtacagccaatggcagttaatatttccgcaaatggttttggatttgctatt 893


Query: 3712 tgtcag 3717
           ||||||
Sbjct: 894  tgtcag 899
```


```
LOCUS       AL590428              163577 bp    DNA     linear   PRI 31-JUL-2001
DEFINITION  Human DNA sequence from clone RP11-553A21 on chromosome 6, complete
            sequence.
ACCESSION   AL590428 AC026605
VERSION     AL590428.7  GI:15072593
KEYWORDS    HTG.
SOURCE      Homo sapiens (human)
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.
REFERENCE   1  (bases 1 to 163577)
  AUTHORS   Chapman,J.
  TITLE     Direct Submission
  JOURNAL   Submitted (31-JUL-2001) Sanger Centre, Hinxton, Cambridgeshire,
            CB10 1SA, UK. E-mail enquiries: humquery@sanger.ac.uk Clone
            requests: clonerequest@sanger.ac.uk
COMMENT     On Aug 1, 2001 this sequence version replaced gi:15021177.
            During sequence assembly data is compared from overlapping clones.
            Where differences are found these are annotated as variations
            together with a note of the overlapping clone name. Note that the
            variation annotation may not be found in the sequence submission
            corresponding to the overlapping clone, as we submit sequences with
            only a small overlap as described above.
            This sequence was finished as follows unless otherwise noted: all
            regions were either double-stranded or sequenced with an alternate
            chemistry or covered by high quality data (i.e., phred quality >=
            30); an attempt was made to resolve all sequencing problems, such
            as compressions and repeats; all regions were covered by at least
            one plasmid subclone or more than one M13 subclone; and the
            assembly was confirmed by restriction digest. The following
            abbreviations are used to associate primary accession numbers given
            in the feature table with their source databases: Em:, EMBL; Sw:,
            SWISSPROT; Tr:, TREMBL; Wp:, WORMPEP; Information on the WORMPEP
            database can be found at
            http://www.sanger.ac.uk/Projects/C_elegans/wormpep This sequence
            was generated from part of bacterial clone contigs of human
            chromosome 6, constructed by the Sanger Centre Chromosome 6 Mapping
            Group.  Further information can be found at
            http://www.sanger.ac.uk/HGP/Chr6
            RP11-553A21 is from the library RPCI-11.2 constructed by the group
            of Pieter de Jong. For further details see
            http://www.chori.org/bacpac/home.htm
            VECTOR: pBACe3.6
            IMPORTANT: This sequence is not the entire insert of clone
            RP11-553A21 It may be shorter because we sequence overlapping
            sections only once, except for a 100 base overlap.
            The true right end of clone RP11-553A21 is at 163577 in this
            sequence. The true left end of clone RP11-525G3 is at 88067 in this
            sequence. The true right end of clone RP3-397H23 is at 2000 in this
```

LOCUS AL591480 91419 bp DNA linear PRI 26-JUL-2001
DEFINITION Human DNA sequence from clone RP11-525G3 on chromosome 6, complete sequence.
ACCESSION AL591480
VERSION AL591480.8 GI:15026959
KEYWORDS HTG.
SOURCE Homo sapiens (human)
ORGANISM Homo sapiens
Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.
REFERENCE 1 (bases 1 to 91419)
AUTHORS Almeida,J.
TITLE Direct Submission
JOURNAL Submitted (26-JUL-2001) Sanger Centre, Hinxton, Cambridgeshire, CB10 1SA, UK. E-mail enquiries: humquery@sanger.ac.uk Clone requests: clonerequest@sanger.ac.uk
COMMENT On Jul 27, 2001 this sequence version replaced gi:14586293.
During sequence assembly data is compared from overlapping clones. Where differences are found these are annotated as variations together with a note of the overlapping clone name. Note that the variation annotation may not be found in the sequence submission corresponding to the overlapping clone, as we submit sequences with only a small overlap as described above.
This sequence was finished as follows unless otherwise noted: all regions were either double-stranded or sequenced with an alternate chemistry or covered by high quality data (i.e., phred quality >= 30); an attempt was made to resolve all sequencing problems, such as compressions and repeats; all regions were covered by at least one plasmid subclone or more than one M13 subclone; and the assembly was confirmed by restriction digest. The following abbreviations are used to associate primary accession numbers given in the feature table with their source databases: Em:, EMBL; Sw:, SWISSPROT; Tr:, TREMBL; Wp:, WORMPEP; Information on the WORMPEP database can be found at http://www.sanger.ac.uk/Projects/C_elegans/wormpep This sequence was generated from part of bacterial clone contigs of human chromosome 6, constructed by the Sanger Centre Chromosome 6 Mapping Group. Further information can be found at http://www.sanger.ac.uk/HGP/Chr6
RP11-525G3 is from the library RPCI-11.2 constructed by the group of Pieter de Jong. For further details see http://www.chori.org/bacpac/home.htm
VECTOR: pBACe3.6
IMPORTANT: This sequence is not the entire insert of clone RP11-525G3 It may be shorter because we sequence overlapping sections only once, except for a 100 base overlap.
The true right end of clone RP11-525G3 is at 91419 in this sequence. The true right end of clone RP11-553A21 is at 2000 in this sequence.